Mining credit card data

Thesis

Mining credit card data

Gordon Blunt, BA

Department of Statistics
Faculty of Mathematics and Computing
The Open University

Submitted for the degree of Doctor of Philosophy
February 2002

AUTHORS NO. LO200280
DATE OF SUBMISSION: 20 FEBRUARY 2002
DATE OF AWARD: 9 APRIL 2002

Abstract

Data mining is the process of finding interesting or valuable structures in large data sets. It is a modern discipline, and takes ideas and methods from statistics, machine learning, data management and other areas. In many ways, it is similar to exploratory data analysis, although the size of current data sets distinguishes between data mining and standard exploratory data analysis. Data mining can pose novel challenges because of the amount of data to be analysed.

This thesis is concerned with modelling different aspects of credit card holders' behaviour and detecting patterns in the ways customers use their card accounts. A review of the literature on consumer purchasing behaviour and data mining is given, and for the latter the differing viewpoints of the statistical and the computer science communities are discussed.

Two types of models are examined: descriptive models, which describe how customers have been observed to behave; and predictive models, which predict how customers are likely to behave in the future. Both types of model are applied to the main types of credit card use: repayment behaviour and transactional behaviour. Relationships between the two different sorts of behaviour are described, and models are examined which can link the two. Some valuable insights and discoveries of potential commercial value are described.

Simple graphical tools are used to illustrate the unearthing of unexpected patterns and relationships, and it is shown how more sophisticated modelling can build on such discoveries.

Acknowledgements

I would like to thank my supervisor, Professor David Hand, for his constant guidance, encouragement and patience. I would also like to thank Drs Niall Adams and Mark Kelly for their help and support.

I am grateful to Barclaycard for allowing me to use data for this work, and partly funding my research; and to Sandra Linney and Alex Kells for helping with data extraction.

Finally, I would like to express my gratitude to Judy, for her patience and tolerance over the years that I have spent working on this thesis.

Chapter 1

1 Introduction and review of previous work

The UK credit card market is in a state of flux. This change is being stimulated by a variety of causes, including new competition, new technologies, increasing customer awareness, and tougher customer requirements. Existing players are faced with problems of attracting new customers in the teeth of greater competition and of retaining customers who may be tempted to move elsewhere. At the other end of the spectrum, card issuers seek to encourage customers to use their cards instead of using alternative payment methods.

Growth has been strong since credit cards first appeared in the United Kingdom (BBA, 2001), although the market was slowing by the early 1990s. In the second half of the decade, with the arrival of new types of competition, credit card growth more than doubled: from less than 7% a year to more than 15% a few years later. Furthermore, there were perhaps 30 different credit cards in circulation in 1990, whereas today's figure is closer to 2,000 (MoneyFacts, 2002).

The impact of these massive changes is described more fully in Chapter 2, but the long established credit card issuers needed to make huge changes to the way they work, the products they provide, and the way they use their data. Some of the data sets are large – in 2000, there were almost 50 million credit cards in circulation, customers spent more than £95 billion on 1.5 billion transactions, and borrowed more than £35 billion (BBA, 2001).

1.1 Aim of this thesis

The aim of this thesis is to model aspects of, and detect patterns in, the way customers use their credit card accounts. In particular, we look at repayment behaviour and transaction behaviour, and use a variety of tools to do so. Some of these come from a statistical background; some from those used by the data mining community.

We will examine two kinds of models: 'descriptive models' and 'predictive models'. Descriptive models describe how customers have been observed to behave. In particular, our aim here is to divide the population of customers into qualitatively distinct behavioural subgroups. In contrast, predictive models predict how customers are likely to behave. Here, we make a further distinction between predictive models which

(1) permit us to choose people who will be most likely to exhibit particular types of behaviour (e.g. often fail to pay their account in full; or the total amount of interest they might pay in their first year),

(2) predict expected future behaviour so that it can be modified (e.g. sometimes, but not often, fail to pay account in full; perhaps with the intention encouraging customers to spend more widely).

Both type 1 and type 2 predictive models are useful for marketing purposes. We call them, respectively, 'first order' and 'second order' models.

Data mining generally deals with messy, distorted data, possibly from samples that have not been properly constructed, and if the formal inferential tools of statistics are

applied, their results may need to be viewed with some caution. One consequence of this is that sophisticated statistical model building tools are likely to be of less relevance in a data mining context. Many of the latter (for example bootstrapping or general linear models) will not result in markedly better models than simpler approaches (e.g. linear regression), yet are likely to take much longer to produce, given the size of the data sets involved. In many ways, data mining has similarities to exploratory data analysis, although the size of current data sets distinguishes between data mining and standard exploratory data analysis, as described by Tukey (1977).

We discuss the distinction between *patterns* and *models* in Section 1.3: a *model* is an overall summary of a set of data, or subset of data; a *pattern* is a small scale feature of a data set. In this work we are mainly seeking to model peoples' behaviour, and that is the reason it is appropriate to use a relatively small (by data mining standards) sample of 10,000 customers and a million transactions. We describe these data more fully in Chapter 3.

1.1.1 Structure of the thesis

After this introduction, the remainder of the chapter reviews previous work, and this is split into two parts – the first describes the modelling of consumers' purchasing behaviour and the second describes data mining. Chapter 2 is a review of the current state of the credit card market in the UK and Chapter 3 gives details of the data set on which our analyses have been undertaken.

In Chapter 4, we describe several ways to define characteristics of repayments to credit card accounts. Customers can repay any amount between the full amount and

the minimum requested by the card issuer (Chapter 2 gives more details). In Chapter 5, we describe a number of ways to predict such behaviour in the future, and in Chapter 6, we estimate the revenue such activities can generate.

Chapter 7 introduces spending behaviour, and we describe characteristics of how people use their cards for spending in different trade sectors. We also describe data quality and data distortion, and explain one sector – petrol stations – in detail, and model some unexpected behaviour in that sector. In Chapter 8 we describe a taxonomy of spending, give parsimonious models for certain characteristics of spending, and describe how different sectors are related. In Chapter 9, we describe how to predict future spending from current data.

In Chapter 10, we link the theme of Chapters 4 to 6 with that of Chapters 7 to 9. In other words, we link repayment and spending behaviour, whereas in the earlier chapters we considered the two aspects of credit card use separately. We show how they are related, which they must be (in some sense) because they are different aspects of customers' use. Often, the relationships between the two types of behaviour may not be immediately obvious.

Finally, in Chapter 11, we give our conclusions, and present directions for future research.

Most of the work that follows describes our attempts to construct models of behaviour, and to find patterns in the data. To do this, we will use a range of techniques, including linear discriminant analysis, *k*-nearest neighbour analysis, principal component analysis, linear and logistic regression, visualisation, graphical models, and association analysis.

1.2 Modelling consumers' purchasing behaviour

1.2.1 Introduction

Much work has been done on aspects of modelling consumers' spending behaviour, particularly the purchasing of packaged grocery brands. There are two main strands to this, one from a statistical, and the other from an econometric, viewpoint. We describe both of these in Section 1.2, and in Chapter 8, we use the methods described in the next section to model the number of transactions in different trade sectors.

1.2.2 The pattern of consumer purchases

There are many ways in which we can describe the spending of credit card holders: for example, the number of transactions made by each customer (in a specified time), or the total amount represented by these transactions, or perhaps the time of year at which they are made. The first and third of these could give us a fourth – that of frequency of purchase, which may vary by customer, by sector and by the time of year. In this section, we describe work others have done to model the number of purchases of packaged goods. This has a close analogue with our data, especially the number of transactions per customer per sector.

1.2.3 Random effects models - univariate parameterisations

Early work described the use of the Negative Binomial Distribution (NBD) to model consumer purchases (Ehrenberg, 1959). This distribution has several characteristics shared by the consumer panel data that Ehrenberg used – it is always positively skewed, has only one mode, often at zero, and leads to a 'reverse J' shaped distribution. Our data show similar characteristics, although typically the tail of our distributions is longer than those shown by Ehrenberg. He used the NBD because it

fitted well the consumer purchasing data sets he examined. An appealing feature was also the relative simplicity of the distribution, because it is completely specified by two parameters. Two further assumptions were made: that the purchases of any particular customer in successive time periods were independent random samples from a Poisson distribution; and that the distribution of the average rates of purchasing μ of different consumers should be proportional to a χ^2 distribution with $2k$ degrees of freedom. Here, k is the exponent of the NBD.

Chatfield et al. (1966) noted another simple distribution that would fit consumers' purchasing data – the Logarithmic Series Distribution (LSD), and realised that in many situations, the LSD and NBD give almost equivalent results. However, for some 'heavily-bought products' Chatfield noted that the fit to the LSD 'is not very good'. We will demonstrate that the NBD suffers a similar problem for our more heavily used sectors, albeit providing a very good fit over most of the data – usually with more than 95% being explained. The NBD model only applies when purchasing behaviour is stationary, and most of our data, over the short period at our disposal, are effectively stationary.

The same authors, in a later paper, (Goodhardt et al., 1984) expanded the model's basic form into the Dirichlet model, which gives a multivariate Beta or Dirichlet distribution across different consumers. It differs from the earlier models, which were for individual brands, but this approach can be used for a product class as well (i.e. multiple brands). Like the NBD, this approach models a stationary market and it introduced some concepts similar to those that will appear in this work. These are as follows.

Penetration is the proportion of the population buying a particular brand, which here becomes the proportion of credit card customers spending in a particular sector with their credit cards. Indeed, Chatfield (1986) refers to 'purchase occasions'.

Repeat buying is the purchase of brand X in subsequent time periods, and here becomes the subsequent use of the same sector.

Purchase frequency per buyer. This is as obvious as the description implies.

The distribution of brand purchases becomes one of several ways that we can use to describe customers' behaviour, such as the number of weeks in which a particular sector is used, or the amount spent in a sector.

Total product purchases becomes total number of transactions made by each customer.

Goodhardt et al. (1984) use the term 'Dirichlet model' as shorthand for the NBD-Dirichlet model, and the NBD is still an important part of the approach, and the authors simplify the calculations to those of a Beta-Binomial distribution. For a specific brand j, with brand share α_j/S in product class S, the probability of making r_j purchases of brand j, conditional on n purchases of the product class having been made ($r_j \leq n$) is given by the Beta-Binomial distribution

$$p(r_j \mid n) = \binom{n}{r_j} \frac{B(\alpha_j + r_j, S - \alpha_j + n - r_j)}{B(\alpha_j, S - \alpha_j)}$$

Where B denotes the Beta function, and $S = \sum \alpha_j$.

Wrigley and Dunn (1985) extended the NBD and Dirichlet models to encompass covariates; Davies and Pickles (1987) to trip timing and store choice of grocery shoppers; while Romaniuk et al. (1999) applied it to the use of products, particularly shampoo. Queen (1999) used a multiregression dynamic model, which followed earlier work on a generalisation of the Dirichlet model to partially segmented markets. Segmented markets can be defined as follows. Suppose a market consists of m brands $\{1, \ldots, m\}$, which are sold to n types of consumer $\{T(1), \ldots, T(n)\}$; associated with $T(i)$ is a type set $C(i) \subseteq \{1, \ldots, m\}$, $1 \leq n \leq n$, containing all of those brands which $T(i)$ chooses. If n = 1, so that there is a single type of purchaser for the whole market: the market is homogeneous. If the type set $\{C(1), \ldots, C(n)\}$ forms a partition of $\{1, \ldots, m\}$, the market is said to be segmented, otherwise it is said to be partially segmented. The analogue in this work is that credit card customers can use their card in a subset of trade sectors, and not use it at all in others.

All of the approaches just described have equivalents in our data, particularly in the number of transactions made, and their frequency.

1.2.4 More complex approaches

Allenby et al. (1998) and Allenby and Rossi (1999) model the heterogeneity in peoples' purchase decisions, primarily by using a Bayesian approach, and conclude that such modelling is preferable to 'classical approaches to modeling heterogeneity' because much of the influence on a particular consumer is 'not well represented by summary statistics'. Their hierarchical approach allows for parameter estimation at what they call the 'disaggregated level', i.e. to smaller groups of consumers. Foekens et al. (1999) use varying parameter models to try to estimate the dynamic effects of promotions, as do Bhargava and Sargan (1984), although the latter applied

their techniques to income data, rather than purchasing behaviour. Dekimpe et al. (1999) used unit root based econometric models to estimate the long and short run effects of price promotions.

Kamakura and Russell (1989) used a latent variable approach to market segmentation, based on the assumption that consumers can be allocated to a small number of segments, 'each characterized by a vector of mean preferences and a single price sensitivity parameter'. Latent class models assume that homogeneous groups of consumers exist, and can be modelled. The authors allowed brand preference and price sensitivity to affect the choice among competing brands.

Chiang et al., (2001) describe weekly purchase data in several packaged grocery product categories, using hazard functions, and concluded that short term promotional activities have little effect on the timing of consumers' purchases. The emphasis of their work is the time between successive purchases, and whether or not that can be influenced by promotional activities. Allenby and Lenk (1994) used logistic normal regression, although they focused on 'brand choice probabilities', which is not really relevant to our work because card holders can buy any product or brand they choose. The analogue for our data is that people may choose, or not, to use their card (or other means of payment) for any particular purchase. In this work, we have little information on the proportion of a consumer's total spending that we see on our credit card, so will not pursue this idea, although a development for future analysis is to look at a customer's credit and debit card transactions when the cards are from the same bank.

1.2.5 The econometric approach

Leszczyc and Bass (1998) review recent models on consumer brand choice and include, in passing, some of the sources we have mentioned. However, their main thrust, along with much of the econometric literature, is to move away from the parsimonious specifications of the NBD approach, which characterises much of the work we described in Section 1.2.3.

Allenby and Ginter (1995) describe a way of segmenting markets, using a Bayesian random effects model, with consumers whose response to particular offers can be defined as 'extreme' in some way. They used customer preferences from a market research study, rather than actual behaviour, for this approach. A possible drawback is that typically 90% of a population lies outside the extremes, but the authors suggest the technique will assist in product design, rather than marketing to the population. They believed that people at the extremes would be more likely to switch to different products, and thus have the most extreme values for product attributes, which could guide new product design. In our data set, use of this type of market research data is not practical, because most of our data are behavioural, and derived from peoples' use of their cards. In many cases, demographic details are not stored, or only updated spasmodically, and are thus of limited use. In the cases where relatively full data are available, it is likely to be among newer customers, but at any one time, new customers typically comprise only around a tenth of the file, which is of little practical use if we seek to model aspects of all customers' behaviour.

Allenby et al. (1999), sought to model 'interpurchase' times of individual customers of an investment brokerage firm by means of a generalised Gamma distribution. A

sentence in the introduction illustrates that perhaps the authors consider the 'traditional' (i.e. more parsimonious) statistical approach to be flawed: 'firms have been forced to use statistical models of purchase data that by necessity ignore competitive effects and unobserved customer behaviour (e.g. purchases of competitive products)'.

1.3 Data mining

1.3.1 Introduction

Data mining has been defined as the 'discovery of interesting, unexpected, or valuable structures in large data sets' (Hand and Blunt, 2001) and as 'the application of algorithms for extracting patterns from data' (Fayyad et al., 1996b). The latter authors, in a later paper (Fayyad et al., 1996c) defined Knowledge Discovery in Databases (KDD) as 'the overall process of discovering useful knowledge from data, and data mining refers to a particular step in that process'. We will use the first definition, rather than the narrower one, because we view data mining as more than the development and use of algorithms. It must involve analysts with both domain and statistical knowledge, otherwise the findings from any data mining exercise could be flawed.

Interest in data mining as a research topic has mushroomed in recent years. A search of the Institute for Scientific Information's 'Science Citation Index' and 'Index to Scientific and Technical Proceedings' found 39 references to papers mentioning 'data mining' in 1996, but that had risen to almost 1,000 by 2001. Many of these are from computer science – Mackinnon and Glick (1999), for example, list a hundred references, half of which are in publications concerned with computer databases.

There are many general introductions to data mining, and we discuss those from statistical and computer science backgrounds in Sections 1.3.3 and 1.3.5 respectively. Some texts are also written for the business user, and concentrate more on potential business applications of data mining, and less on methodology. Two recent examples are Berry and Linoff (2000) and Rud (2001).

We draw a distinction between *patterns* and *models*, and use the definition of Hand, Blunt, Kelly and Adams (2000). A *model* is an overall summary of a set of data, or subset of data, so it is thus the 'standard' statistical use of the term. It could be, for example, a linear regression model, or a conditional independence graph, or one of many others. In this work we are mainly seeking to model peoples' behaviour, and that is the reason it is appropriate to use a relatively small (by data mining standards) sample just over 10,000 customers. In another sense, ours is not a small data set, because there were more than 1,000,000 transactions in the two years for which we had data. We describe these data more fully in Chapter 3.

A *pattern* is a local structure, referring to a small number of objects, where 'small' is context dependent. An example in our case could be the 0.5% of petrol station transactions that are for amounts of more than £100 (the 99^{th} centile of this distribution is just over £50). We describe transactions in this sector in some detail in Chapter 7. Another could be the 0.2% of customers who maintain a 'negative balance' for a whole year. We describe our approach to dealing with these customers in Chapter 4.

1.3.2 Data mining and visualisation

An important aspect of any data analysis is visualising the data. The power of visualisation methods derives from the ability of the human eye to perceive patterns, which is what it evolved to do. We will describe, in many parts of this work, different ways of looking at data, and the importance of using appropriate methods. In this instance 'appropriate' means that the structure or interesting or useful aspects of the data become apparent. We will also give examples of how difficult it can be to select the most suitable methods, especially with large data sets, and show that simple techniques, such as scatter plots or histograms, may be suitable, providing they are used in a sensitive way. In general, some sensitivity and awareness of the properties of the methods are needed in order to avoid mistaken conclusions. Examples in Chapter 8 show that truncated histograms can reveal structure that is completely concealed in histograms showing all of the data. We also show how an increase in median spend per sector, as the number of sectors increases, conceals the fact that the majority of customers had relatively little spending. We will also show how, if used insensitively, the analyses themselves can conceal structures.

Hand and Blunt (2001), said 'Visualisation is very important in data mining, and some highly sophisticated tools, which can process vast datasets and which are based on supercomputers have been developed. Of course, visualisation in huge data sets is but the natural development of older graphical methods, many of which are still effective in quite large data sets – though sometimes care has to be exercised. [...] Perhaps we should also note the advantage of graphical and visualisation tools in helping to convince (possibly non-numerate) senior management of the reality of the discoveries that have been made.'

Others who have written at length about the importance of visualisation are Tufte (1983), Chambers et al. (1983) and Cleveland, (1993). It was also an important part of works such as Keim and Kriegel (1997), Cox et al. (1997) and Mackinnon and Glick, (1999). One of our aims in this thesis is to illustrate how these 'standard' (and, indeed, classical) graphical methods can be applied in data mining.

1.3.3 Data mining from a statistical perspective

Many authors have written on this subject, for example Chatfield (1995), Elder and Pregibon (1996), Glymour et al. (1996, 1997), Huber (1997), Hand (1998a), Mackinnon and Glick (1999), Hand (2000b), Smyth (2000) and Hand, Mannila and Smyth (2001). The last of these is an interdisciplinary text, and could appear in our review from a computing, as well as a statistical, perspective. Several themes run through these sources, as follows.

Parzen (1997), in his consideration of the relationship between the statistical sciences and data mining, defines 'core statistical research' to be 'about mathematically synthesising ideas drawn from many analogous applications'. He proposed the use of an alternative name – 'statistical methods mining', and suggested, like Fayyad et al. (1996c) that data mining cannot take place without statistics.

(1) Size of the data sets

Many huge data sets exist today that contain with terabytes of data. For example, we refer below to the SKICAT analysis of Fayyad et al. (1996a) that has three terabytes of astronomical data. Hand, Blunt, Kelly and Adams (2000) give several examples, such as the 7 billion transactions a year made in Wal-Mart stores, and Pregibon (2000) describes the 70 million telephone long distance calls handled a *day* by

AT&T. Statisticians have not generally confronted data sets with millions of cases, and tens of thousands of variables, mainly because such data sets are relatively recent.

Several problems arise as a direct consequence of data sets this large, especially when we need to analyse them, and Wegman (1995), Glymour et al. (1996, 1997), Huber (1997) describe some of the problems that can arise. Hypothesis tests may not work, because they can lead to the rejection of models no matter how closely they seem to fit the data. We give examples of this in Chapters 4 and 7, on repayment behaviour and petrol station transactions respectively. Traditionally, statisticians have used α values of 0.05 or 0.01, and these might be wholly inappropriate with massive data sets (Glymour et al. (1997) say that 'this point is of fundamental importance for data miners').

Huber (1997) describes the fact that standard errors and other tests of significance might not make sense when applied to 'samples of convenience' or 'opportunistic' data sets. Many data sets in the data mining literature appear to be of this nature, as pointed out by Glymour et al. (1997) and Hand and Blunt (2001). Greenfield (1994) gave examples of researchers – in disciplines other than data mining – who experienced problems because they collected data without sufficient thought about how the objectives of their research could be formulated in terms of a statistical problem. In data mining, similar problems are likely to arise because data are often taken as an opportunistic sample from a large data set and presented to the analyst. She or he may have no influence on how the data are derived, or how the sample is constructed.

Parameter estimation can also be problematic (Glymour et al., 1997). Classical statistics treats the parameters as fixed but unknown, and Bayesian statistics provides a framework for estimating the distributions of parameters, when the data under investigation come from a sample that is governed by probability distributions. These assumptions are often violated in data mining contexts. Also, as Hand (2000b) pointed out, the data miner's main concern may lie elsewhere, especially with the speed of the algorithm used to produce a model. There may well be an emphasis on sequential algorithms where only a single pass through the data is required.

(2) Data mining can be pattern focused rather than model focused

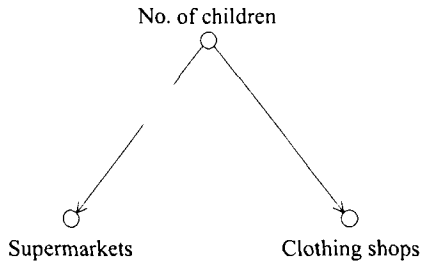
This relies on the definition of *patterns* as local, often small, structures (Hand, Blunt, Kelly and Adams, 2000) or local dependencies (Glymour et al., 1997). This is distinct from *models*, which are overall summaries of the entire data set (Glymour et al., 1997, Hand, Blunt, Kelly and Adams, 2000). Data mining, and the methods used in a particular application, are driven by the needs of the application itself. If patterns are of importance (for example in plastic card fraud detection) then there might be no alternative but to search for all occurrences of particular patterns in the data. In the case of credit cards, fraudulent transactions occur very infrequently: 0.17% of credit card turnover in the UK in 2000 (BBA, 2001), but it is the industry's objective to prevent every such transaction, if possible.

(3) Problems with causal inference

As noted by Glymour et al. (1996), 'causal inference from uncontrolled convenience samples is liable to many sources of error'. Statisticians are well aware of this – the fact that A and B may be correlated does not mean that one causes the other.

Problems arise from three main areas: latent variables, sample selection bias, and model equivalence.

Latent variables are unrecorded features that affect variables that are recorded in the database, and whose variation influences some of the recorded variables. The result can be an association between two variables that is not a result of the direct association between the features themselves. For example, we might find that our data indicate a relationship between the amount a customer spends in two particular trade sectors. However, a possible explanation for this relationship would be that shown in the following graphical model, where spending in the two sectors is conditionally independent given the number of children. Our data set does not contain this last variable, so it would not be possible to identify that the correlation was spurious.



Sample selection bias can occur when data are not a properly specified sample (see the penultimate paragraph in (1) above). Whether selection bias is important is dependent on the objectives of the analysis – if we seek to make inferences to an underlying population, any sample bias can invalidate our results. For example, to select a 10% sample, it is easy to select every 10^{th} case, starting from a random point in the first ten. However, if there is some relationship between successive members then there is a danger that there could be some periodic structure in the list. The data

we describe in Chapter 3 were not affected by this, but shortly after we extracted our sample there were several 'upgrade' programmes. Customers were given different cards, depending on their behaviour. Tranches of account numbers were allocated sequentially because of these programmes, and had we taken our sample later than we did, we would have had to have taken note of this in the sampling method we used.

Model equivalence occurs when it is possible to find different models that all adequately fit the data, and which could have 'quite distinct causal implications' (Glymour et al. 1996, 1997). A procedure that arbitrarily selects one (or a small number) of these models can lead to different inferences. Huber (1997) also notes that models can be distorted by selection bias.

One data mining technique outside 'traditional' statistics may be prone to confusion between correlation and causality: that of association rules, where the notation itself could be the source of misunderstanding. Association rules were defined in Hand, Blunt and Bolton (2001) as follows. Let V be a set of attributes, corresponding to the set of all possible items that can exist in a database. A *transaction*, T, is a subset of V. An *association rule*, R, is a pair (A, B), where A is a subset of V, and B is (usually) a single element of V, not in A. A is often called the *antecedent* of the rule, and B the *consequent*. A *transaction*, T, is then said to *satisfy* a rule $R = (A, B)$ if the elements in R are all in T. For any C a subset of V, P(C) represents the proportion of transactions that include C as a subset. That is, $P(C) = \mathit{frequency}(T \mid C \subseteq T)$. Association rules are frequently shown as $A \Rightarrow B$, and sometimes described as an 'implication' (see, for example, Agrawal and Srikant (1994), Dong and Li (1997),

Pasquier et al. (1999), Berry and Linoff, (2000)). We discuss association rules more fully in Chapter 8.

(4) Contaminated data

We described sample bias in (1) above, but there is another form of distortion where individual records are corrupt because of missing values or incorrectly recorded values (Hand, 2000a, Hand and Blunt, 2001). Any modelling done without regard to these missing or distorted data is likely to be seriously flawed, and thus any inferential statements which follow from them (Hand 2000a, Glymour et al., 1997). Similarly, the use of any automatic pattern detection methods are likely to have problems if much data distortion is present.

(5) Non-stationarity and spurious relationships

Other problems can arise too, such as non-stationarity of the data (Hand 1998a, Mackinnon and Glick, 1999). Analyses conducted at different times might not be consistent as populations change over time, a phenomenon sometimes known as 'population drift'. In the case of the work we describe, most of our data are from 1996 and 1997, and the credit card market has evolved rapidly in the years since, as has Barclaycard. Any analyses we were to undertake on more recent data would need to take account of this evolution.

(6) Scalability

We mentioned in (1) above that speed of algorithm is often important, especially with a massive data set. Researchers who develop data mining tools might focus on the computing problems they face, of which speed is only one. This can lead to rejection of some of the newer, iterative, statistical techniques which do not 'scale up' to large data sets. Wegman (1995) describes the impact on computing times of

some techniques. Even in the relatively short time since he wrote this paper, the analysis of 'huge' (10^{12} bytes in his classification) data sets is not unknown, as described later (Fayyad et al., 1996a, with their three terabyte data set).

Mackinnon and Glick (1999) describe how other aspects of the analysis of huge databases are affected as well. They explain trade-offs between the (potentially) conflicting requirements of data storage and retrieval, and how they are discussed in the database literature. The amount of information that can be represented on a visual display, and how the human eye can perceive it, also affects visualisation, which we discuss next.

(7) Difficulties in visualisation

We contend that data analysis, if done properly, has always involved visual inspection of the data. When data sets were 'small' (a few hundred cases and a few tens of variables), it was possible to examine everything. However, as Wegman (1995) and Huber (1997) point out, human abilities in pattern recognition break down with massive data sets. The former also notes that traditional methods of graphical data analysis do not always scale to 'large and huge' data sets.

To make matters worse, some data reduction algorithms (clustering or discriminant analysis, for example) may not scale well either. This could leave the analyst with no option but to take a simple random sample, with the risk that one may lose some of the structure one is seeking (see (2) above). Keim and Kriegel's work (1996) is devoted solely to the visualisation of large data sets for data mining. They describe five types of technique – pixel oriented, geometric projection, icon based,

hierarchical and graph based, and apply their system to three real data sources and a simulated one.

(8) Simpson's paradox

Examples of data 'lying' are given in Elder and Pregibon (1996) and Glymour et al. (1997), although they are mostly illustrations of Simpson's paradox (Simpson, 1951). Most statisticians will be aware of the impact of interactions in such analyses, but data miners, perhaps with little formal statistical training, may not be. Add to this the impact of massive databases and the consequent need to find fast algorithms, and there is potential for misinterpreting results quite badly. Hand (1994) gave several examples of incorrect analyses that arose because the wrong questions were asked.

Bradley et al. (1999) noted that the 'query formulation problem' has not received much attention in database research. They pose the question about how the computing community 'provide access to the data' when users do not know how to describe their objectives in a specific query. We suggest that this is an area for more study, and it is one where statisticians ought to be able to help the data miners. This is because the queries might be easy for a person to formulate, but difficult to code as a Structured Query Language (SQL) query. For example, 'is this credit card transaction fraudulent?' Questions which are mechanistic, on the other hand, are easily coded – for example 'tell me the mean spend of all credit card holders on their Gold card in the Tyne-Tees television region'. Hand (2000b) asserts that statistics has four important activities - *data exploration*, *description*, *modelling*, and *inference*, but that data miners are concerned mainly with the first two of these, so the relative lack of development of structured queries should not be surprising.

1.3.4 Data mining and 'data dredging'

Chatfield's paper (Chatfield, 1995) is written from a more 'traditional' statistician's viewpoint than much of the work we have described thus far, and much of the discussion about data mining is disparaging – he uses the term pejoratively some of the time. He describes the circumstances under which data mining is less than rigorous from the statistician's viewpoint. For example, 'the situation where the analyst looks at a new set of data with virtually no preconceived ideas at all. The rather derogatory terms *data mining* and *data dredging* are sometimes used in this context to describe procedures of [this] type, particularly when the analyst eschews careful thought based on external knowledge in favour of deriving the best possible fit from a large number of entertained models.' Most of the other work we have described is concerned with the relationship between statistics and data mining, and the synergy that needs to coexist at the interface between the two disciplines, and the necessity of working with large data sets.

Chatfield's paper describes the importance of the correct formulation and selection of models, rather than parameter estimation for existing ones, which is nearer to the typical data mining scenario. Perhaps our final comment on the traditional statistician's viewpoint should be from Wegman (1995): 'data sets of large and huge size require intellectual attention. If statisticians do not pay attention to these problems, other scientists will'. In a complementary vein, Friedman (1997) quotes Efron as saying 'those who ignore statistics are condemned to reinvent it'.

1.3.5 Data mining from a computing perspective

There are several good introductions to data mining, notably Mannila (1996), Fayyad et al. (1996b, 1996c), Piatetsky-Shapiro et al. (1994), Witten and Frank (2000) and

Han and Kamber (2001). We included Hand, Mannila and Smyth (2001) in Section 1.3.3, but could equally well have shown it here too, because it is an interdisciplinary text. Most acknowledge that statistical rigour is a crucial part of the KDD process. Fayyad et al. (1996c) say 'Knowledge discovery from data is fundamentally a statistical endeavor. Statistics provides a language and framework for quantifying the uncertainty that results when one tries to infer general patterns from a particular sample of an overall population.' Mannila (1996) notes that statisticians have often been derogatory about data mining (as we noted in Section 1.3.4), and refers to work by Tukey (1977) that describes *exploratory data analysis* (EDA). Much of our work is EDA, and we would agree with both of these authors, because we recognise the importance of using the data to guide our analyses. Fayyad et al. (1996b) say that two goals of data mining 'tend to be *prediction* and *description*' and later that 'description tends to be more important than prediction' in data mining, which is similar to the EDA referred to by Mannila (1996). We now summarise some of the important points about KDD and data mining.

(1) Huge data sets are now available

The author of many papers in this area, Usama Fayyad, has worked on several such data sets himself, and we describe such an application later. Piatetsky-Shapiro et al. (1994), describe, in their review of KDD 93, applications of data mining in science, finance and manufacturing, such as an 'extremely large' supermarket database (we mentioned Wal-Mart in a similar context), and the analysis of tropical storm data.

Not only are they already large, in some instances, but Bradley et al. (1999) note that the growth rate of data sets is far exceeding any manual capabilities for analysis of those data. Mannila (1996) notes that much KDD work is done on huge data sets,

but that KDD methods 'can be useful even on small data collections'. In the latter case, some might not consider it data mining if applied to small data sets (Smyth, 2000), when it might be more appropriate to use more traditional statistical methods.

(2) Confusion between patterns and models

We have already defined patterns and models, but references in some of the data mining literature are rather confused. Bradley et al. (1999) made the distinction between patterns and models, and gave the following definition: 'a pattern is classically defined to be a parsimonious description of a subset of data. A model is typically a description of the entire data'. This is almost the same as our definition in Section 1.3.3 (2).

In the introduction to their paper, Fayyad et al. (1996b) described KDD as being in need of tools 'to *intelligently* and *automatically* assist humans in analyzing the mountains of data for nuggets of useful knowledge'. The use of the word 'nugget' suggests that the original focus of the early workers in the area was pattern detection rather than model building. Other workers have failed to make the distinction. For example, Ha and Park (1998) describe 'discovery-driven data mining' as a 'bottom up approach that starts with the data and tries to get it to show something new'. They then go on to describe a variety of common data mining techniques, some of which are usually used to model the data (e.g. cluster analysis) and some which are typically used for pattern detection (e.g. association rules). The authors do not make a distinction between modelling the data and finding patterns in it, as we did in Section 1.3.

(3) Interestingness

This is a theme that recurs in the data mining literature, and Fayyad et al. (1996b) refer to it as 'an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity'. This is context dependent, however, as the authors say: 'this definition of knowledge is by no means absolute. In fact, it is purely user-oriented, and determined by whatever functions and thresholds the user chooses'. Bradley et al. (1999) discuss evaluation by several criteria, which are almost the same as those given by the earlier authors – validity, utility, novelty, understandability, interestingness.

Other authors strive to find ways to automate the process of finding interesting items (see, for example, Freitas (1999), or Padmanabhan and Tuzhilin (1999). The first of these strives to find what we might class as 'departures from the norm', while the second uses a very similar concept, that of 'unexpectedness'. Both of these terms are context dependent, as illustrated by the following statement, that they 'use a user-specified support threshold value to determine if the subset [of interest] is large enough'.

In our view, Fayyad et al. (1996b) are correct when they say 'there should be a champion who can define a proper interestingness measure e.g. a domain expert who can define a proper interestingness measure *for that domain*' (our italics). The implication is that the quest for an automated assessment of interestingness is doomed to failure.

(4) The KDD process

Most authors agree that data mining is an iterative *process* (see, for example, Fayyad et al., 1996b, Mannila, 1996, Klemettinen et al., 1997, Bradley et al., 1999, Hand

and Blunt, 2001), which cycles backwards and forwards between the miner and the domain expert (who could be the same person). It involves the 'repeated application of specific data mining methods or algorithms and the interpretation of patterns generated by these algorithms' (Fayyad et al., 1996b), and we agree that this could be fraught with difficulties if it is done without awareness of the statistical issues. Fayyad et al. (1996a) and Bradley et al. (1999) call this 'the blind application of data mining methods', which they note 'can be a dangerous activity'.

The KDD process involves many steps, and can include pre-processing the data, sampling from the entire data set, or projecting to a smaller number of dimensions, before the data mining step (Fayyad et al., 1996b). All of these need to be done carefully, and if the pre-processing involves the removal of outliers, then, as Glymour et al. (1997) point out, the analysis of truncated data can lead to problems.

Bradley et al. (1999) and Mannila (1996), echo Fayyad et al. (1996b) in saying that data mining is one step in the KDD process, which can be messy, with many iterations over previous steps. After each step, it can become apparent that the data should have been pre-processed in a different way, or that previously unforeseen patterns exist, and so cause the analyst to look for slightly different patterns. Indeed, the work on petrol station transactions in Chapter 7 came about almost entirely in this way, after we found a surprising (to the author at least) number of transactions at exact pound amounts, with higher peaks at multiples of £5.

Fayyad et al. (1996b) give guidelines for a successful KDD application, which they describe as follows.

- The impact that the analysis will have, which for a business could be measured

as increased revenue or lower costs.

- That no good alternative methods exist.
- Sufficient data should be available: the more complex the task, the more cases are needed. Also, that this should be relevant information, with as few errors as possible. We discuss this further in Chapter 3.
- Prior knowledge (such as that possessed by a domain expert) is 'most important'.

Many authors (for example Fayyad et al., 1996b, Mannila, 1996, Hand and Blunt, 2001) agree that domain knowledge is essential to the data mining process, a judgement with which we would agree wholeheartedly. Heckerman (1997), comments that anyone who has performed any 'real-world' modelling knows the importance of domain knowledge.

(5) Caveats about KDD

We mentioned in (4) that 'blind application of data mining methods [...] can be a dangerous activity', because 'invalid patterns can be discovered without proper interpretation' (Fayyad et al. 1996b, and Bradley et al. 1999) . Fayyad et al. (1996c) also claimed that 'knowledge discovery from databases is fundamentally a statistical endeavor', with which we agree, and it is similar to the view we expressed earlier in this chapter.

Piatetsky-Shapiro et al. (1994) describe problems that can arise if the analysis is carried out with insufficient statistical awareness, as do Glymour et al. (1997). Many of these are a direct result of the large sample sizes that are common in data mining. We have already mentioned that significance tests carried out with an α level of 0.05 or 0.01 might be wholly inappropriate if the data contains thousands of cases. Glymour et al. (1997) describe an example where a model has parameters that are

'highly significantly different from zero, even when the training data are pure white noise'. We give examples of a similar phenomenon in Chapters 4 and 7, where we model repayment behaviour and petrol station transactions respectively. In our case, the models fit the data very well by eye, but the predicted values are significantly different from the observed values.

Problems can arise when we try to visualise large data sets (Bradley et al., 1999), because human abilities do not scale up to massive volumes of data. We have already mentioned (in Section 1.3.2) the power of the human eye and brain to discern patterns, but mentioned that diagrams need to be drawn with sensitivity to the needs of the analysis in hand. Projections to lower dimensions can transform the problem into a much simpler, possibly linear one, but finding the optimum projection can itself be difficult task. Finally, Mannila (1996) noted 'as the area is a mixture of techniques from different fields, there is also the danger of reinventing old (bad) solutions'.

(6) Computing issues

An important issue facing computer scientists and those from the database and machine learning communities is the sheer size of the data sets at their disposal. As soon as a data set cannot fit into main memory, the time needed for analysis will rise by orders of magnitude, because accessing data from a disc could be a million times slower than from main memory (Smyth, 2000). As Hand and Blunt (2001) note, this has led to an emphasis on sequential algorithms, where only a single pass through a data set is required.

Data mining has drawn from several disciplines, such as statistics, database management, pattern recognition, artificial intelligence, optimisation, visualisation,

high performance and parallel computing (Bradley et al., 1999). This in turn has led to a new set of terms, such as data warehouse, OLAP (online analytical processing), drill down, roll up, data cube and so on. As Mannila (1996) describes, the number of variables in large data sets is the root cause of complexity, because combinations grow exponentially. In some of the analyses we describe later, with only 26 variables, the number of combinations is $2^{26} \approx 6.7 \times 10^7$, and this is a relatively small data set compared to, say, that of Goodall (1999) in Section 1.3.6, with 10,000 variables, which has an astronomically large number of combinations.

1.3.6 Examples of data mining

(1) Astronomical classification

Fayyad et al. (1996a) describe their 'sky imaging and cataloguing tool (SKICAT)' which they had developed a couple of years earlier. Their task was to classify 3,000 digital images of 23,040 × 23,040 sixteen bit pixels each, resulting in more than 3 terabytes of data. The classification was based on training examples provided by the user, and they used a machine learning approach to construct their classifier. Image processing software was used to produce attributes that described each of these objects, and 40 of these were measured automatically. An important part of their approach was that they projected the high dimensional pixel space onto a lower dimensional feature space, which then allowed them to transform the problem into one that was solvable by a supervised learning algorithm. This sort of data reduction was referred to in Fayyad et al. (1996b, c), and it can be important to reduce the number of variables under consideration, as the number of possible combinations of variables rises exponentially (Mannila, 1996, Bradley et al., 1999). The authors were successful in their task, and 'exceeded their initial accuracy target of 90%'. To

validate the performance of their classifier, they drew test sets of data independently from the training data, and then performed statistical tests (although they don't say which) to measure the rate of conflicting classifications.

(2) Atmospheric data analysis

Macedo et al. (2000) use simple graphical tools to explore multivariate atmospheric science data which has a spatial component. Their approach is similar to some that we will describe later – that the use of simple graphical methods can be illuminating, and that interesting patterns can be 'easier to spot'. Their data are from an ocean atmosphere data set, dating back 150 years. The core of their approach to visualisation is the linking of data in many views. For example, they use linked brushing to extract information about the conditional distributions of multivariate data (see Cleveland, 1993, for a simple introduction to brushing). Some of our analyses use this technique to identify cases that might be outliers, or causing some distortion in plots. They give, as one example, how they could analyse the relationships between sea surface temperature, sea level pressure, wind speed and direction over time.

(3) Telephone fraud calling patterns

Cox et al. (1997) also stress the importance of visualisation, and use visual data mining techniques to detect telephone calling fraud for AT&T's network. Their approach is based on the premise that 'human pattern recognition skills are remarkable', and they list several other applications where their methods have been applied successfully. There are parallels between fraudulent call patterns and fraudulent use of credit cards – for both, fraud is a small proportion of overall transactions, but the overall cost is significant – they estimate $1 billion in the U. S.

Their visualisation technique allows users to see unusual patterns quickly, and they claim that it complements automatic data mining techniques in four ways.

Firstly, people excel at detecting patterns, but tire easily when faced with routine repetitive tasks. Secondly, fraud is dynamic, and fixed threshold algorithms are easily detected and 'bandits' are able to outwit them. Thirdly, domain knowledge is crucial, because a telephone network's best customers have similar calling patterns to those of fraudsters – they make many expensive calls. This is similar for credit cards – some of the best customers use their cards extensively, and again this might be a typical fraudulent pattern. In both cases, the aim is to withdraw the service from fraudulent users, but to encourage legitimate customers to make calls as much as possible, and terminating the service incorrectly would be quite likely to result in the loss of a valuable customer. Thus it is important that a person be involved. Finally, the visualisations are available immediately, and the user has complete control over the dimensions that are displayed, which can complement traditional data mining tools.

The last two applications are examples of *pattern* finding, as distinct from *model* building, a distinction we have already made. If our objective is to search for patterns, we might need to examine all cases in the database: it may not be enough to use a sample. As mentioned before, a pattern is a local structure, and in the case of fraud detection, could refer to only one example that displayed particular characteristics.

(4) Health care delivery

Goodall (1999) describes a problem that is almost quintessential data mining – it is *secondary* (Hand, 1998a) or *retrospective* (Glymour et al., 1997) because the data

were originally collected for purposes other than analysis, and it has a 'large p problem' with more than 10,000 variables. As the author says, 'questions of accuracy and specificity – the appropriateness of the data to address particular clinical and financial concerns – predominate over statistical issues of sampling, estimation and even censoring of data'. Traditionally health care data are used to assess the clinical effectiveness of particular treatments, but Goodall's objective is to analyse the health care system, seen as an aggregate. Analysis of large health care data sets can help by identifying where savings can be made, and resources better directed. There are a large number of variables: for example, he cites more than 10,000 diagnosis and procedure codes, which can be used singly, in aggregates or combinations, which could be used in regressions. These are then present for millions, or tens of millions, of patients.

Goodall stresses the importance of correct presentation of data, which he describes in terms of an analogy with a subway map for London or New York, which are characterised by their simplicity, but still have a dense information content. He explains the statistical challenges that are presented by health care data, many of which are very similar to those of credit cards. Each patient (in our case, customer) is an individual, and records are not interchangeable. No single set of variables can capture all the information about an individual: the analogy for credit cards is that we do not know what other mechanisms they use for spending or borrowing (e.g. debit cards, cash or overdrafts). Groups of interest in health care have relatively low frequencies; one example he cites has 131,323 patients at risk, but only 17,076, or 13%, had the disease. This proportion is almost exactly the same as one we describe in Chapters 4 and 5, which is concerned with partial repayers. His conclusion is that

these characteristics 'throw the burden of analysis towards exploratory techniques'. He then suggests some techniques which could be used with health care data, but describes three factors which are important – statistical computation (and computational performance); the need to understand the data, where and how it originates; and the need to organise computations depending on the data and the desired outcome.

(5) Study of a particular type of crime

Adderley and Musgrove (2000) use a neural network (a Kohonen self organising map, later referred to as SOM) to classify offenders in a particular sort of crime, and to try to link different incidents to the correct offender(s). The authors describe how the keeping of computerised records as 'control information for management' also allows analysis of the data too: 'this has led to the development of the area of Computing known as Data Mining', and they refer to the SOM as a 'specific data mining technique'. We agree with the second point, but not the first, which we discussed in Sections 1.3.3, 1.3.4 and 1.3.5. The performance of their method was assessed by a police officer who was not part of the original research team, and this was necessary, as the authors did not have a test set to validate their results.

The authors seem to have an unusual idea of validation: 'few of the crimes have been successfully detected (i.e. solved) and hence there is *no perfect solution to act as a comparison*' (our italics). Statisticians realise the futility of searching for a perfect solution. In the event, they had a police sergeant check their results, and gave him more information than they used in their modelling. He reported that some members were 'in his opinion clearly different to the majority of members of the cell' and that some crimes were ascribed to people appearing in 'widely differing' cells. Data

quality problems, and the difficulties in coding statement details because of witnesses' recall of the same person varied dramatically, and meant that any statistical technique might have had problems.

(6) Risk analysis and targeted marketing

Jha and Hui (1998) describe the credit scoring of data from a large bank, and illustrate the use of several data mining tools to provide risk analysis and targeted marketing. They say 'association rules provide a good understanding of the model', but we will show (in Chapter 8), that – at least on our data set – they do not in themselves, and we would cite Padmanabhan and Tuzhilin (1999) who describe a study that resulted in 20,000 rules. We would argue that this leads to another problem – the analyst's capacity to assess such a large number of rules. Also, this is only a small subset of the possible number of rules in most data mining data sets. The data we will describe have a possible 2^{26} rules, and if there are m variables, there are 2^m possible rules. Hoffman and Wilhelm (2001) present a new measure, the *difference of confidences*, and use graphical methods for 'getting an overview of a set of association rules'. We will explain this more fully in Chapter 8, and describe how we used similar techniques, and why they did not add a great deal of insight with our data.

(7) Paleoecology

Mannila et al. (1998) compare machine learning, exploratory rule mining, and some statistical methods applied to paleoecological reconstruction, where the objective is to determine environmental temperatures in the past. The statistical methods they consider are different variants of regression (e.g. linear and inverse linear) and the

construction of a Bayesian model; the machine learning methods presented are *k*-nearest neighbour and regression tree learning.

The authors say that all of the methods have their strengths and weaknesses, and it is difficult from their presentation to reach a conclusion about the superiority of any particular method. For example, the 6-nearest neighbour method has the lowest cross validation accuracy, but its variability is the largest of the techniques they have shown. This makes it difficult to discern any trend of rising temperatures over the last 5,000 years, which has been shown by the 'results from other data sets'. In contrast, one of the regression techniques and the Bayes model, showed a slight increase in temperatures over the period.

(8) Credit card fraud detection

Chan et al. (1999) develop a method for detecting credit card fraud, using 500,000 credit card transactions from two American banks, and each data set includes 15% - 20% of fraudulent transactions. They describe how their approach can cope with skewed distributions (by partitioning the data into subsets which have the 'desired distribution', applying mining techniques to these subsets, then combining the classifiers). These desired distributions are determined by 'extensive sampling experiments'. However, the 'distributions' explained in the paper do not appear to be the same as they would be if described by a statistician. They refer to the relative proportions (i.e. the 20% and 80%) of fraudulent and legitimate transactions, and the results they describe are highly dependent on these proportions. The closer the proportions approach 0%, the worse the authors' algorithms (mostly tree based) performed. In the UK, 0.17% of credit card transactions are fraudulent (BBA, 2001), so the authors' claim to perform better than commercial software might need closer

investigation if applied to a data set similar to the one we describe in Chapter 3. For tasks where the target group is of the order of 20% of the total, their methods perform well.

(9) A major bank and a large marketing application

Hunzicker et al. (1998) describe data mining to be an integral part of deploying a loyalty based customer management programme, not something to be done for 'the sake of its own beauty'. They describe some elementary statistical analysis (calculating means, correlations between variables), but perhaps misunderstand some statistical concepts with the comment that 'all balanced data sets must still have reasonable size (more than 10,000 records) to *guarantee* statistical significance' (our italics). See Glymour et al. (1997) for some comments on the size of data sets, and their impact on statistical analysis.

The models described were all tree based, and as part of the building process, the authors found that the time consuming part of the process was not the model building itself, but the discussion with domain experts in the bank. Earlier in this chapter we argued, as others have done (e.g. Fayyad et al. 1996b) that the domain expert is an essential part of the data mining 'team', and time can be saved in the long run by having statisticians and domain experts involved at all stages.

1.3.7 Data mining tools

As Hand and Blunt (2001) noted: 'Many tools are used for data mining, some of which (such as cluster analysis) have been in existence for some time, but others of which have come to prominence with the power of modern computers and massive

data sets (for example association analysis). The older tools typically have their roots in statistics, and the later ones in machine learning.'

We could divide these into two broad classes – those that seek structure in data sets in which no variable(s) has been singled out as special (e.g. as response variables) and those such as regression where the aim is to build models which enable one to predict a single (numerical) response variable from one or more predictor (or explanatory) variables.

Many other tools exist for prediction (e.g. linear discriminant analysis, and nearest neighbour methods which we use in Chapter 5), and some more computationally intensive tools, such as neural networks, projection pursuit regression or generalised additive models may have limited use in large data sets because of the time required to estimate the parameters. Much depends on whether the aims can be achieved by means of an analysis based on a sample of the data.

Bradley et al. (1999) divide data mining tools into five classes: predictive modelling, clustering, dependency modelling, data summarisation and change and deviation detection. The methods they describe are regression (linear and non linear), classification, clustering, summarisation, dependency modelling, change and deviation detection, decision trees, example based methods (e.g. neural networks), probabilistic graphical models, relational learning. They go further, and provide illustrations of programming for data mining methods, and illustrate these with case studies. Fayyad et al. (1996b) give a similar breakdown.

To illustrate the problem of 'scalability', which we mentioned in Section 1.3.3 (6), consider cluster analysis. Bradley et al. (1999) give algorithms for performing

cluster analysis, and the methods they describe, like other authors in data mining, are based on pre-selecting k clusters. This is natural, because a hierarchical clustering needs a distance matrix to be computed, and this will have n^2 elements (where n is the number of cases), and this could be impractical. In Barclaycard's case, if we were to use the whole data set of approximately 10^7 cards, we would need a matrix containing at least 10^{14} elements.

Time series models (see, for example, Box et al. (1994), Fuller, (1996) or Harvey, (1989)) may be used in data mining and can be applied to various aspects of credit card behaviour, but they are not a central feature of this work. Han and Kamber (2001), and Hand, Mannila and Smyth (2001) give examples of time series analysis in data mining (they cover the related issue of sequence retrieval as well). Much of the methodology they describe is concerned with global models of the whole data set, but, as we described in Section 1.1, our main aim is to model aspects of the ways that individuals use their credit card accounts.

Customers' behaviour at an individual level is irregular, but can be made more systematic if we aggregate customers into groups, similar to the phenomenon described by Goodhardt et al. (1984) in the context of consumer purchasing behaviour. We will mention seasonality, but only briefly, in Chapter 7, and also give examples of the variability between peoples' spending patterns. In the first case, we simply look at total spending per sector, and how each month differs from the average. In the second, we make no attempt to model individuals' behaviour for trend, seasonality or any of the other 'usual' elements of time series analysis.

Examination of seasonal patterns at an individual level may be possible, but would require longer runs of data than we describe here. We would also need to carry out much exploratory work to deduce the level of aggregation that might be required to ensure reasonably regular behaviour in each group. However, deducing the 'correct' amount of aggregation – to result in a series that can be modelled by the usual time series methods – is not trivial. Also, at least three years of data (probably five) is likely to be necessary if we want to be confident about being able to identify temporal components. Barclaycard routinely stores three years of account history data, but only 15 months of individual transaction data (see Chapter 3 for a fuller description of these terms). The reason for this is quite simple – the size of the data set grows massively each month, with another 30 million transactions and all of the associated fields that comprise each record.

Sufficient accuracy can be achieved by using simple summary measures of large groups of customers' behaviour (e.g. amount spent per month per sector) because the business is often seeking to model the whole file, or large parts of it. Work done by Barclaycard indicates that such 'whole file activity' is seasonal, with predictable peaks and troughs, and an identifiable trend. A range of univariate and multivariate techniques is used extensively to model and predict a variety of different business indicators (e.g. spending, the number of telephone calls received per month). The main thrust of our work, however, is the exploration and prediction of individual behaviour, rather than the construction of models to deduce trend (or seasonality, or whatever) across the whole business.

1.4 Summary

We have described a broad sweep of previous work, encompassing consumer purchasing, data mining and statistics, and the interface between the latter two. We discussed data mining from the viewpoint of workers from statistical science and computer science, and gave examples from wide range of applications to which data mining processes have been applied. We covered, but only briefly, some more complex approaches, because our data are not suitable for the techniques discussed by the authors in those fields.

We described some problems that are particular to data mining, and difficulties that can arise when trying to use statistical methods on datasets typically found in data mining. Some of these problems will feature in this work, such as contaminated data, extremely skewed distributions, the necessity of using secondary data, and problems with visualisation. We have avoided issues of scalability by using a fairly small (in data mining terms) sample, although it was still time consuming to use particular techniques, such as *k*-nearest neighbour methods, because of the necessity of computing distances between all pairs of customers.

Chapter 2

2 The history of the UK credit card market

2.1 Introduction

This chapter describes the state of the UK credit card market, the sorts of changes that are taking place, and why those changes are occurring. The remainder of this section sets the scene. Section 2.1.1 briefly describes what credit cards are, comparing them with the other major plastic cards, and Section 2.1.2 summarises how the credit card market works.

Section 2.2 presents a review of how the credit card market has reached its current state. What is apparent, from the graphs and figures shown there, is how quickly things are changing. In some cases, the rate of growth seems to be exponential, raising natural concerns about what might be to come should there be a recession.

Recent years have witnessed particular competition from two new sources: one is American card companies expanding into the UK, with a strong home base and considerable experience to draw upon, albeit in a slightly different market. The other is UK consumer organisations expanding into the credit card market, driven by pressures in their own industrial sectors. Such competition presents new challenges for the established players. Whatever they may do, one thing is certain: existing credit card companies cannot remain unresponsive to the challenges posed by these aggressive and determined new sources of competition. Section 2.2.3 examines these new entrants into the market.

The entry of new players changes the shape of the industry. Instead of a simple choice between a handful of cards, the consumer is now faced with a choice between a wide variety of cards, each with slightly different properties, opportunities, rewards, and attributes. The nature of the industry, and not merely the nature of the competition, is changing. Section 2.2.4 examines the growth of multiple card ownership.

Finally, Section 2.3 draws some conclusions and examines what is just around the corner. In order to cope with the increasing competition (and the increasing determination of the competition) the industry is inevitably becoming more mathematicised: more sophisticated mathematical and statistical models must be used if market edge is to be retained. Moreover, new technologies such as smart cards, data fusion and customer value models are beginning to be developed.

2.1.1 Credit cards and other cards

Credit cards are 'buy now, pay later' products. The card holder will typically receive a statement once a month and will then have to make a repayment by around the 25th day (known as the 'payment due date') after the statement. The card holder can choose to repay any amount between the total on the statement, and some specified minimum. This minimum amount will have been agreed at the time that the customer received the card. If the holder chooses to make only a partial repayment, then she or he will pay interest on the remaining balance. Credit card issuers all belong to one of several associations, the biggest two in the UK being Visa and MasterCard. Any credit card issued under one of these schemes can be used at any merchant that accepts Visa or MasterCard cards. There are more than 600,000 (BBA, 2001) such merchants in the UK, and millions around the world. A

credit card may be the only relationship between the issuer and the customer - it need not be related to any other financial service. Indeed, most new entrants do not offer a full banking service, and their operating costs are thus likely to be different from the established traditional bank issuers.

In 2000 there were some 50 million credit cards in issue in the UK, representing total spend of £95.2 billion, an average transaction value of £62; an average spend of more than £1,900 per card (British Banker's Association (BBA, 2001)). The BBA figures, though, also include Visa and MasterCard charge cards. Other card issuers in the UK are American Express, Diners, and some retailers issue credit cards too. The latter are of two types; one example is Marks and Spencer, which runs its own financial services arm. The other is the Arcadia type (Burtons, Dorothy Perkins, Evans, Principles and Top Shop), which is run by a third party; GE Capital in this case.

Credit cards are not the only type of plastic payment card. Others include charge cards, corporate cards for business to business purchases, retail (or store) cards, and debit cards. A common characteristic is that they can all be used at a merchant for the purchase of goods or services.

Charge cards differ from credit cards in that the card holder must repay the full amount which is shown on the statement. *Retail cards* are similar to credit cards in that they can be used to pay for goods and services, and have the repayments spread over time. However, they are restricted to one particular retailer or group of retailers, and cannot be used more widely. *Debit cards* are not free standing products, but are available with some current accounts. When purchasing goods or services, they operate in exactly the same way as a credit or charge card, but the amount of the

purchase is deducted from a customer's current account, usually within one or two days. There is no credit facility, per se, although some current accounts will have an agreed overdraft facility.

The number of debit cards in issue has risen steadily, with almost 50 million now being in circulation. In 2000 £76 billion was spent on them (APACS, 2001b), with an average transaction size of just over £32.50. Figures from the Credit Card Research Group (CCRG, 2001) indicate that a far higher proportion (about 30%) of debit card spending is in the 'food and drink' sector than it is for credit cards (where it is about 11%). Conversely, credit cards see a higher proportion of spending in sectors with higher transaction sizes such as travel (13% compared to debit cards' 7%). The average debit card spend in 2000 was around £1,600 per card (APACS, 2001b)

2.1.2 The credit card market cycle

Figure 2.1 shows the main elements called into play when a credit card transaction occurs. We begin at the bottom of the figure, when the card holder purchases some goods or services from a *merchant*. The general term 'merchant' is used to convey the wide nature of credit card spending: about half of this spending takes place in retail outlets, and the rest in bodies such as hotels, travel agents, and petrol stations.

The merchant has a relationship with an *acquirer*, which is an organisation (often a bank) that takes a merchant's transactions, and reimburses the merchant for the value of each transaction. Depending on the agreement that a merchant has made with its acquirer, this will usually be within a day or two of the transaction being made. The *merchant service charge* is the amount that the acquirer charges the merchant for

providing this service. In the UK, it averages around 1.5%. Each merchant has an agreed *floor limit*. Transactions below this amount go through without authorisation, but transactions above this amount need to be authorised. The merchant contacts its acquirer (often by a telephone call), which then checks with the *card issuer* (see below) to see if the transaction can be accepted. Since this authorisation occurs at the point of sale, it must be done quickly. It is now possible to have it done electronically, by the latest terminals, as they can read a customer's details from the card's magnetic stripe or computer 'chip'. Merchants are guaranteed payment on all transactions below the floor limit and all authorised ones above this limit.

The acquirer forwards the transaction to the *association*, most often Visa or MasterCard in the UK, which acts as a clearing house, taking in transactions from various acquirers and passing them to the appropriate card issuer.

There is a fee payable to the association for each transaction, and some transactions do not pass through the associations at all. They are usually those for which a customer uses his or her card in an outlet which (coincidentally) has its transactions acquired by the card holder's bank. The customer will usually not know of the relationship before the transaction, and it will almost certainly have no bearing on the desire to use that particular outlet. Such transactions are kept entirely within the bank's own systems, and thus incur no association fees.

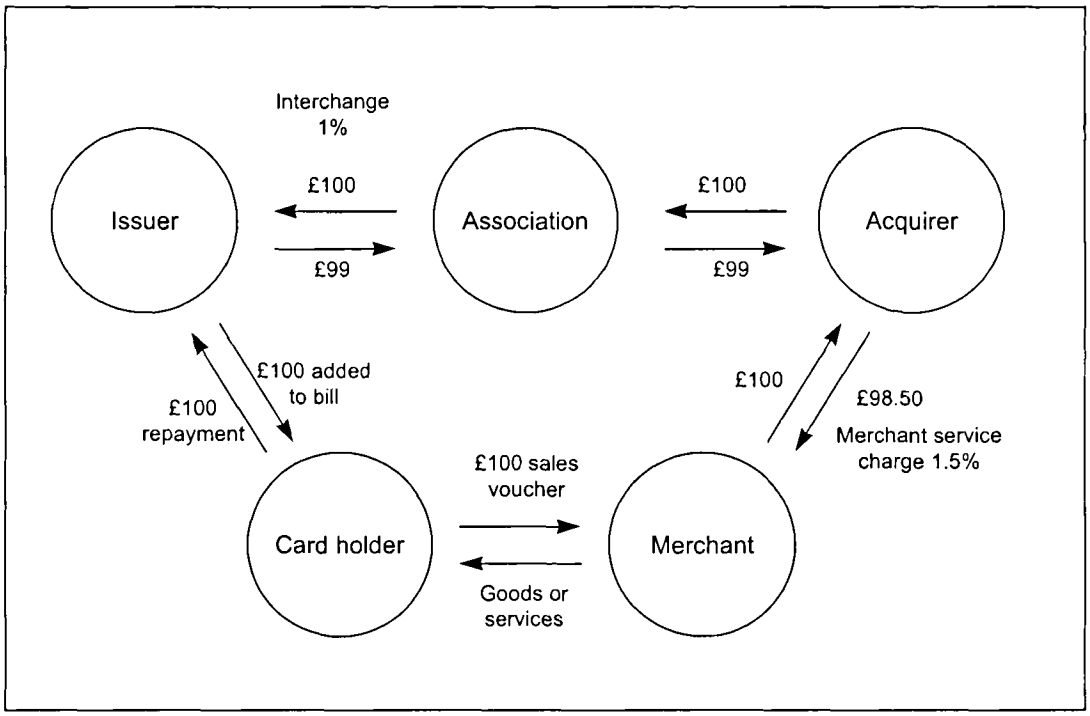
The issuer then sends the amount of the transaction, less the *interchange* amount, to the association. Interchange pays the issuer for carrying the value of the transaction between the time they pay the association and the time the card holder pays them, fraud losses and bad debt. The issuer applies the full amount of the transaction to the card holder's statement at the same time as paying the association, but will not

receive payment until after the next statement is sent out and the card holder actually pays - which may be up to eight weeks later. Interchange is usually around 1%, although it varies depending on the type of transaction.

In March 2000 Don Cruickshank produced a report on competition in UK banking for the Chancellor of the Exchequer (Cruickshank, 2000). The report notes how the credit card market has changed, and hints at the speed of change - 'prices in this product segment move fast: rates below 5 per cent were available a few months after this analysis was undertaken' (late in 1999). It also notes that charging on credit cards and the benefits offered by issuers have changed in recent years.

Most of the major impact (on credit cards) from the review is likely to come from reform of interchange, which has a whole appendix devoted to it. This concluded with the comment 'There is a strong case for reform of the interchange fee system. None of the arguments against change is convincing.' The review estimated that the credit card issuing industry earns around £650 million from interchange, but that fraud losses and cost of providing an interest free period for customers (together) accounted for half of this amount. It seems likely, therefore, that there will be some legislation which could cap interchange, and that could have a significant effect on issuers' profitability, especially those with 'rich' loyalty schemes (for example 'cash back' offers). Some of these offer a rebate as high as 1% of the amount spent. At the moment, much of the cost of funding this rebate will come from interchange, which is itself just over 1%.

Figure 2.1 The main elements of a credit card transaction.



2.2 The development of the UK credit card market

Barclays Bank launched the first UK credit card in 1966. Barclaycard promised retailers it would publish the name and address of every shop which had agreed to take the card at the launch. These appeared in the *Daily Mail* of 29th June 1966, extended over eight pages, and showed 30,000 names and addresses (and took 200 staff from Barclays all night to proof read). Some six years later, in 1972, a consortium of banks originally comprised of National Westminster, Midland (now HSBC), Lloyds (now LloydsTSB) and later joined by Williams and Glyns and the Royal Bank of Scotland group, launched the Access card. In 1977 the international Visa payments system was founded and the rate of introduction of new cards increased: in 1978 TSB launched Trustcard, in 1980 American Express launched its

Gold Card in conjunction with Lloyd's Bank, in 1981 the Co-operative Bank launched its Visa card, in 1982 the Bank of Scotland launched its Visa card and the Barclays Premier Card was launched, and so on. In 1988, following the deregulation of financial services, Building Societies entered the credit card market. In 1993, the first American credit card issuers entered the UK market, and in 1996, UK supermarkets launched their own credit cards. We shall say more about these very recent developments and their likely impact on the future of the UK credit card market below.

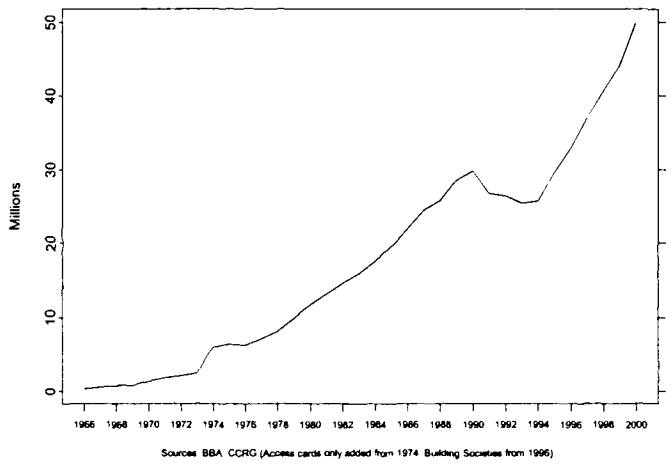
We give a graphical illustration of how dramatically things are changing in Figure 2.2 (see also Crook et al., 1994). In 1989 the Monopolies and Mergers Commission (MMC, 1989) noted that in the two years it had taken to compile its report, ten new credit card issuers had entered the market, and the change since then has been even more dramatic, especially since 1993. Brooking (1997) noted that there were more than 800 different credit cards issued by 1996. MoneyFacts (2002), an agency that provides a tracking service to the financial services sector, estimates that, taking into account all of the different affinity products on offer, there are now more than 2,000 credit cards available. Market Movements (AC Nielsen MMS, 2001) estimates that, where there were around 50 different credit card mailings a month in 1995, this figure had risen to 200 in some months of 2000.

As well as sheer numbers, the complexity of the products on offer has also increased over time. The Monopolies and Mergers Commission report (MMC, 1989) noted a wide range of credit card Annual Percentage Rates (APRs), ranging between 15.4% and 26.8%, with the major issuers around 26.0%, but nowadays there is a far more complicated mix of APRs, fees and charges, with differences in calculation methods

making comparisons difficult. Rates available in the UK ranged between 0% and almost 40% at the time of writing. There can be many 'hidden' costs, which need not be included in the advertised APR, for example: fees charged when a payment is not made by a due date, and charging if a replacement statement or a copy of a transaction voucher is requested (see also Ritzer, 1995).

Figure 2.2 shows a roughly linear growth between the early 1970s and the end of the 1990s, followed by the first fall ever in 1991. This reduction in card numbers followed the introduction of fees by Barclaycard, Lloyds (now LloydsTSB), HSBC (then called Midland Bank), and National Westminster. (The fall was also coincident with a period of recession, but an earlier period of recession, in the 1980s, had no discernible effect.) The rate of increase since 1994 has been even more dramatic than the earlier period. It is probable that this reflects increased competition, as new entrants such as US providers and other sectors diversify into financial services. We shall discuss this further in Section 2.2.3.

Figure 2.2 The number of Visa and MasterCard cards in issue each year.



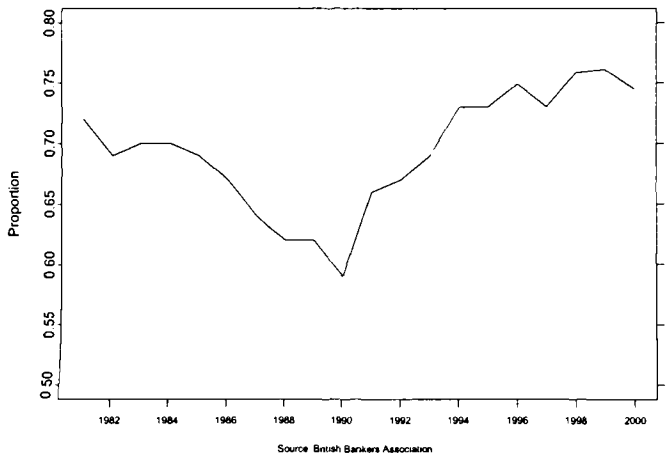
2.2.1 Credit card profitability

Auriemma and Coley (1992) noted that in the US 'during the 1980s, bankcards emerged as one of the most profitable services offered by banks'. Several authors have noted that card issuers in the US produce high profits by keeping their interest rates much higher than their cost of funding customers' balances (Stavins, 1991; Park, 1997; Brito and Hartley, 1995; Ausubel, 1997). These authors describe how credit card interest rates have remained high relative to base rates, and how consumers still choose credit cards, rather than cheaper forms of borrowing. In this country, Brooking (1997) notes that 'credit cards are assets with "sticky" rates', meaning that they tend not to follow the fluctuations of base rates very closely. Apart from this, there appears to be little which has been written in recent years describing the UK experience.

UK banks have rarely made public the proportion of their income that comes from interest paid on credit cards. The Monopolies and Mergers Commission (MMC, 1989) reported that around 70% of the major banks' credit card income was from this source. However, as Figure 2.3 shows, the proportion of balances that incurred interest fell to the lowest level ever seen in the years up to 1990. With customers repaying a higher proportion of their balances, and the interest-earning part of total balances being a smaller proportion than ever, the issuers effectively had to fund an extra £1.1 billion, from which they were not receiving any interest income. Worse, at the same time base rates were high (14%), so that the cost to the issuers was high. In order to cope with this, and to try to develop other income streams and reduce dependency on interest income from 'revolving' balances, fees were introduced. Most long term rates, as opposed to short term promotional ones, are around 13% –

18%. This suggests that credit card companies in the UK have been successful at keeping their rates high - and, by implication their profits too. We have already noted that Cruickshank (2000) came to a similar conclusion, although he did point out that spreads had declined. The implication of this is that although credit cards are a profitable part of a retail banking operation, they are less so than they might have been if competition had not changed the market from 1993 onwards.

Figure 2.3 Proportion of credit card balances which paid interest



Recently, there has been widespread use of low interest 'teaser' rates, often for only a limited period. These are usually well below the longer term rates, and have been used extensively in America. They are a distinguishing feature of the new competitors in the UK market (since 1993) and will be discussed further in Section 2.2.3.

2.2.2 Growth in credit card activity

Figure 2.4 shows the amount spent per card; Figure 2.5 shows the total amount outstanding (i.e. owed) on credit cards at the end of each year. The dramatic growth

is apparent, especially in most recent years. Concern about this will be discussed in Section 2.5. Likewise, concern has also been expressed about the growth in the proportion of total consumer credit that is being taken on credit cards - shown in Figure 2.6.

Figure 2.4 Annual spending per credit card (not adjusted for inflation).

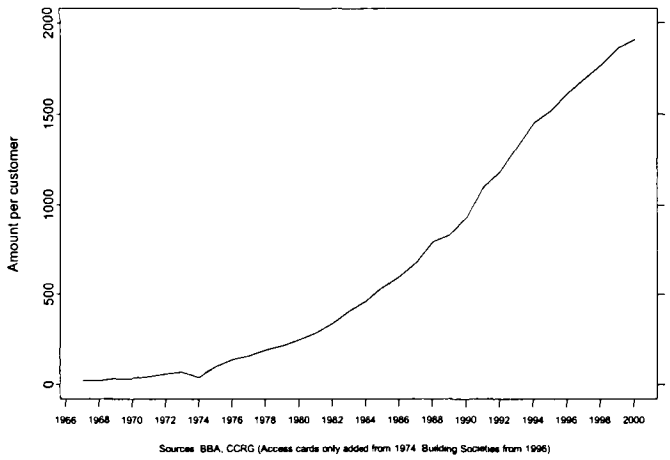


Figure 2.5 Amount outstanding on credit cards, by year (not adjusted for inflation).

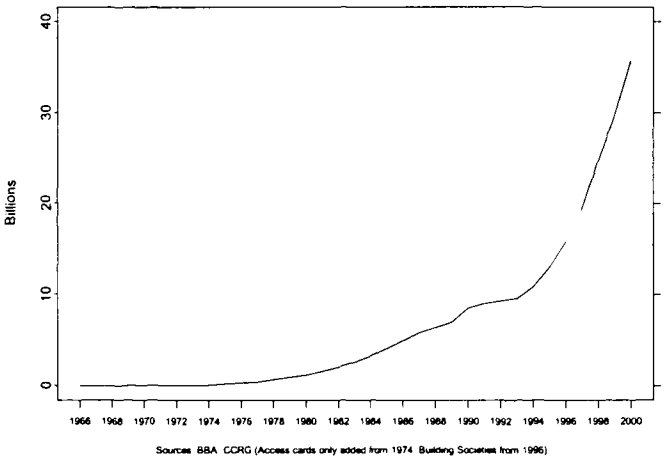
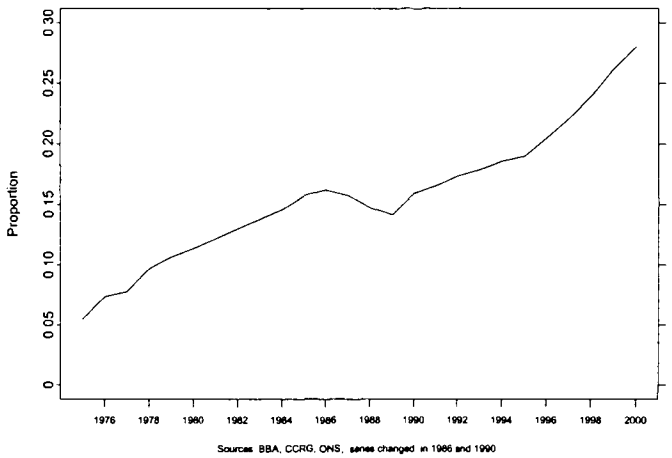


Figure 2.6 Proportion of total consumer credit taken by credit cards.



2.2.3 Recent changes in the UK credit card market

We saw, in Section 2.2.2 and Figure 2.5, how the rate of growth in balances on UK credit cards had increased dramatically since 1994. This increase was primarily due to two new types of competitor entering the market, with high levels of marketing activity: American card issuers already successful in the USA, and strong UK consumer brands facing increasing competition in their own areas and hoping to use credit cards to generate customer loyalty to their core business. Issuers in the first category include MBNA, HFC with its GM branded card, Capital One, People's Bank, and Bank One. Issuers in the second category include the banks that issue cards for organisations like Goldfish (linked to British Gas), Tesco, Sainsbury, and Virgin. For the customer, however, the issuer is of secondary importance for these cards, and it is the loyalty brand that is important.

The first new entrants arrived around the turn of the year 1993/4 – the GM Card and MBNA – which were later followed by Advanta, Capital One, People's Bank, Associates and, more recently, Providian and Bank One. To illustrate the speed of

change in the market, Advanta (at first in a joint venture with the Royal Bank of Scotland) is now wholly owned by RBS; People's Bank and Associates have been bought by Citigroup; Providian has almost collapsed in the USA, and Bank One is owned by Halifax/Bank of Scotland.

The latter half of 1999 saw another new tranche of competitors, whose offering was based solely on servicing over the internet. They were of two types – *egg*, which is solely an internet bank, and *marbles* and *smile*, which are the internet brands of, respectively, HFC and the Co-operative Bank.

MBNA and Capital One are often referred to as 'monolines', and are the largest such issuers in the US. This term originated in America and characterises companies whose business is focused in just one area. In the case of MBNA and Capital One, focusing on credit cards, it means they have no (usually expensive) branch banking network or (often unprofitable) current account systems to operate, and concentrate almost exclusively on high profit, low cost operations. Credit cards are attractive here because these can easily be operated in isolation from other financial products.

MBNA's strategy is based on engendering loyalty and appears to be successful. It provides credit cards for 4,700 'affinity groups' such as the American College of Surgeons and the National Society of Professional Engineers. Reichheld (1993) states that 'at MBNA, a 5% increase in retention grows the company's profits by 60% by the fifth year.' Its profit in 2000 was more than $1.3 billion, which was 28% more than in 1999, which was itself 32% higher than 1998. In the UK, according to MoneyFacts (2002), MBNA has more than 800 affinity partners, and it has more than 4,500 in total. Its 2000 Annual Report claimed that it had 12% of the credit card market in the UK and Ireland.

Capital One, on the other hand, uses what it calls an 'information based strategy', as this quote from its 2000 *Form 10K* illustrates. *Form 10K* is a financial report that publicly quoted companies in the US must file with the Securities and Exchange Commission, Washington DC, every year.)

> *'IBS is the cornerstone of our marketing strategy, and since its introduction in 1988 we have steadily increased our marketing efforts.*
>
> *'An important element of our risk management process is our use of IBS to identify and target potential consumers. We attempt to continuously monitor and improve the effectiveness of our information systems and processes such that senior management possesses the tools to manage portfolio risk and respond to changing market conditions and challenges.'*

One important characteristic of the issuers described above, and Providian a little earlier, is that they are not the major American banks. Their success has been built on the monoline strategy, their emphasis on customer retention, and also on the policy of targeting clearly defined groups of customers. Often these customers already have a credit card, so that their repayment record and credit history are established. This permits the issuers to target customers who will bring in substantial revenue at low risk: a 'good' customer is one who repays a small proportion of the amount outstanding, but does not default.

'Teaser' interest rates, usually much lower than the standard rate, are often used to attract new customers. Of course, these are almost invariably for a limited period - perhaps five or six months - and often only for balance transfers from one card to another. Stavins (1996) has examined the propensity of consumers to borrow on

credit cards irrespective of the interest rates. She found that people are sensitive to rates, but those who are most aware are those with repayment problems. The lowest introductory rate at the time of writing is 0%. Such low rates are often hedged with caveats. In the US, teaser rates have been criticised for enticing 'unwary customers into applying for a card without realizing that they soon will be paying a higher rate,' (Ritzer, 1995).

Ritzer (1995) attributes the very high level of credit card debt in the US to the huge amount of advertising by the industry. He remarks that the industry has made 'credit card debt so easy and attractive that many of us have become deeply and perpetually indebted to the credit card firms.' The difference in debt between the American companies and the UK market total (figures are not available for individual UK issuers) is shown in Table 2.1, which shows the ratio of end of year balance to total year's card spending. The American companies tend to have values greater than the UK market total, sometimes substantially so. The top four rows show four of the largest US companies, and the second block of four rows shows four bank issuers. Providian's figures indicate that its products are used more as a line of credit, than being used for day to day spending. The UK market figures are starting to approach the lowest American issuers, which reflects the preponderance of balance transfer activities that have been such a feature of the market in recent years.

Table 2.1 Ratio of end of year balance to total year's spend.

	1996	1997	1998	1999	2000
MBNA	0.76	0.70	0.69	0.65	0.66
Bank One	0.64	0.92	0.70	0.49	0.47
Providian	0.84	0.90	0.97	0.92	1.02
Capital One	0.87	0.69	0.61	0.66	0.62
Citigroup, US cards	0.49	0.47	0.50	0.46	0.48
Chase	0.78	0.82	0.69	0.62	0.60
Household	0.54	0.51	0.45	0.50	0.46
Bank of America	0.51	0.48	0.39	0.42	0.42
UK market total	0.30	0.31	0.34	0.36	0.37

Source: The Nilson Report (www.nilsonreport.com), the British Bankers' Association (BBA, 2001). The figures are shown as they were at the time, without any attempt to back calculate what each issuer's figures would have been excluding mergers and acquisition activity. Figures refer to credit card activity. This means that some numbers have changed, as some issuers used to include total outstandings in their reporting.

In addition to targeting potentially profitable customers and then working to retain them, the issuers can also encourage people to use their cards. The quotation from the *Form 10K* of Capital One above indicates that a wide range of incentive schemes are being explored. Feinberg (1986) and Ganzach and Karsahi (1995) describe some examples of different kinds of credit card marketing activity aimed at encouraging customers to use their cards.

The traditional issuers have responded to these incursions into the marketplace in various ways. Co-branded offers, in which the issuer collaborates with another organisation and in which both partners benefit, have been set up. For example, Barclaycard is collaborating with Green Flag and Legal & General and NatWest with Air Miles. Similarly, the traditional issuers have also established relationships with new British entries to the card market. Finally, they are also competing on price, though perhaps not as aggressively as the American entries. Examples include Barclaycard's variety of offers for balance transfers, NatWest's introductory rates of 5.9% to new customers, HSBC's 11.9% the Royal Bank of Scotland's 3.9% and RBS Advanta's 1.9% (MoneyFacts, 2002).

The second recent major change that the UK credit card market has seen has been the entry of UK companies with powerful consumer brand images. These include the credit cards of some of the supermarket chains. Points are awarded when the cards are used, with typically double the points when they are used in the parent supermarkets themselves. The card holder can then redeem the points for discounts. These schemes, though, unless the data are well used, can be costly. For example, Safeway withdrew its loyalty card in 2000, and Sainsbury's was reported to be having problems (Mail on Sunday, 11th June 2000). At the time of writing, of the supermarkets, only Tesco had invested much in its marketing of credit cards. If the others did decide to invest heavily in them, they could have a big impact on the market since their potential customer base is huge. Research from the British Market Research Bureau's Target Group Index survey (British Market Research Bureau, 2001) shows that each of them has around 10 million shoppers. (To put this into perspective, Barclaycard, the biggest card brand in the UK and Europe, has 11 million cards in circulation.) With the purchase of Asda by WalMart, an American retailer which is well known for its data mining abilities (see, for example, Hand, Blunt, Kelly and Adams, 2000), the whole area of supermarket loyalty marketing could have its profile raised. However, an interesting dilemma arises. It is likely that some of the supermarkets' shoppers would not be eligible for a credit card if scored by standard methods (Hand and Henley, 1997), so that their applications would be rejected. This would hardly dispose them kindly towards shopping in the supermarket in the future.

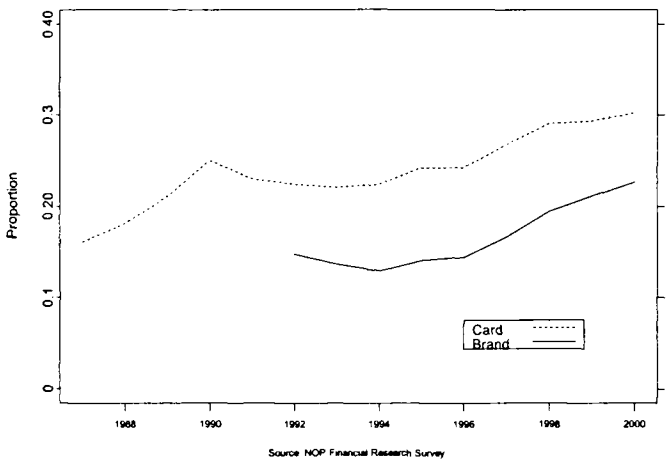
The supermarkets also stand to gain substantially in terms of the extra information they gather about their customers. If a customer uses cash or some other card, then

the store knows that certain items have been purchased together but cannot associate these with a particular individual. Use of the store's own card means that this missing link can be made, permitting much more powerful data mining explorations (Hand, 1998a; Hand, Mannila and Smyth, 2001). In particular, it means that incentivisation schemes can be tightly targeted. DunnHumby (2001) claims that analysis for Tesco's Clubcard scheme has resulted in coupon redemption rates doubling to '20-40%', cost per redemption falling by 50% and customer complaints have reduced by 75%.

2.2.4 Multiple card ownership

Figure 2.7 shows the proportion of credit card holders who hold more than one card, and the proportion who hold cards from more than one issuer, which we refer to as different brands of card. The trend is an increasing one, with multiple card ownership having roughly doubled over ten years.

Figure 2.7 Multiple card and multiple brand ownership.



In some cases, the distribution of multiple ownership will come as no surprise. It is partly a consequence of the policy, by the new entrants, of targeting existing card holders. One possible explanation is that an existing card holder who has already responded to an approach by a relatively older 'new' market entry (such as MBNA) is also likely to respond to an approach by another, possibly newer, issuer (Egg, for example). A natural market segmentation seems to be developing, with various implications for issuers. In particular, it suggests that the new entries may face special problems in retaining their customers.

2.3 Conclusions: the future

The consumer credit industry is becoming increasingly mathematicised, as is demonstrated, for example, by Hand and Henley (1997) and Thomas (1998). This is a response to the increasingly competitive nature of the business, to more stringent customer requirements, and to the opportunities offered by modern computing storage and power. The industry is following the development of the financial markets which, commencing about a quarter of a century ago (Black and Scholes, 1973), began to develop more sophisticated mathematical and statistical models to handle the risk which is the essence of the industry, although the nature of the mathematical and statistical models being used differ. The quotation given above, from the *Form 10K* of Capital One, conveys the flavour of the increased sophistication of the models being adopted.

Over the next few years we expect to see substantial efforts made to learn more about individual customers via record linkage, data fusion, and data mining activities, to develop holistic customer models (using, for example, Bayesian belief networks and graphical models: (Hand et al., 1997; Stanghellini et al., 1999), and to develop

customer value models. The analysis of customer behaviour at individual transaction level has barely been attempted, but it promises to yield an immense amount of information - and the technology is now in place for this to be analysed. Data files of billions of entries (the British Banker's Association (BBA, 2001) gives the number of credit card transactions in the UK in 2000 as 1.5 billion) can now be manipulated and explored.

Examination of transaction patterns will permit more effective market segmentation. This is already undertaken to some extent. For example, Gold cards are given to customers with higher incomes and, beyond this, there are Platinum and Titanium cards (the latter in the US). A different kind of segmentation is being explored by American Express - their *American Express Blue* credit card is aimed at the under 35s. Egg and the other internet based cards are aimed at people who are happy to conduct their whole relationship with a credit card issuer on line. We have already noted a quite different basis for modelling customer behaviour in Section 2.2.3 - how people respond to approaches from new issuers. It is clear that there are many opportunities for refining customer models and developing new products finely tuned to customer requirements.

At a lower level, new card technologies are being explored in parallel with the above developments. Kemp et al. (1997) provide a detailed examination of the use of photographs on cards. Boxall (1996) reports that these are popular with customers, but they may not be very effective in preventing fraud: it is expecting a lot to expect a possibly harassed shop assistant to scrutinise carefully a picture around a quarter of the size of a postage stamp. Even in controlled conditions they are not a reliable way

of allowing legitimate transactions while preventing fraudulent ones (Kemp et al., 1997).

A more promising development is the use of the computer chip on a card. Not only can this advance be used to combat fraud, but it can also be used to store and control all other details of a customer's financial transactions, including current account, savings account, mortgage, and so on. Of course, merely because this is possible does not mean that people will want to do it: there is clear evidence of the 'jam-jarring' phenomenon, in which people keep different financial services with separate companies or products.

The payments market, and the organisations in it, have become more fragmented in recent years. This is likely to continue, especially with the advent of new technologies such as internet commerce and the continued growth of telephone banking.

Section 2.2.3 examined new entrants to the UK market. There is no reason to suppose that things will stop there. Indeed, there are some obvious potential entrants:

- Microsoft has just launched a credit card, in conjunction with First USA.
- British Telecom is in a unique position in the UK because it holds payment records on around 22 million households - and the best predictor of future behaviour is often past behaviour. (Although one must tread warily: AT & T ran into problems with its credit card - for example, the net profit on the 'AT & T Universal Card Services' portfolio fell from $93 million in the first 9 months of 1996 to $27 million in the first 9 months of 1997 (Figures from the 3rd Quarter Form *10-Q* for AT & T, 1997). AT & T has since sold this card portfolio to Citibank.)

- We have already noted that British Gas launched the Goldfish card, in an attempt to encourage customer retention. As the utilities industry suffers increasing customer attrition in the wake of deregulation, we may expect more efforts of this kind.
- 'e-commerce' could change the payments market fundamentally. At present, there are many estimates of the scale of this impact (Governments and research companies are producing their own estimates of the effect of this). Credit cards seem to be the ideal way to pay, and some issuers are advertising that they guarantee security on electronic purchases. Also, a report by Europay (Europay press release, 6th October 1999, Waterloo) estimated that 75% of web purchases are by credit card. However, the impact of mobile telephony, interactive TV, and technologies yet to emerge could all dislodge credit cards from their pre-eminent position.

The UK credit card industry is in a state of change. New technologies, new customer requirements, new opportunities, increased competition, both from the home environment and foreign competitors, are all serving to stimulate progress. The profile of the industry seems set to evolve dramatically over the first decade of the third millennium.

Chapter 3

3 Barclaycard credit card data

3.1 Introduction

All of our data are drawn from the Barclaycard Visa credit card database. We had an extract of a representative sample of 10,212 customers, who had accounts with the company in July 1997, and we received their data from January 1996 to July 1997. Later, we received transactions for these customers covering the remaining months of 1997. This subsequent data set was extracted a few months into 1999, and the vast majority of customers were the same, but there were some differences – see Section 3.2.1 for a fuller explanation of these. This illustrates a consequence of one of the differences between data mining and 'traditional' statistics – a distinction we made in Chapter 1. Data mining is often secondary data analysis; the data may be sampled opportunistically, and not drawn from a properly constructed sample.

We described some very large data sets in Chapter 1, and drew the distinction between *patterns* and *models*. Our aim in this work is primarily concerned with modelling rather than pattern detection, so it is legitimate to look at a representative sample, rather than the entire customer database. We also noted that data mining generally deals with messy, distorted data, and that the formal inferential tools of statistics may need to be used with caution.

Most of the data are derived from customer behaviour, although there is some limited demographic information, and all of the records are based on a customer's 'billing month'. To ensure smooth work loads, around 5% of statements are produced on

each working day of every month, and any particular customer's statement will always be printed on the same billing day, regardless of the date on which that day falls. To illustrate, billing day 1 is usually the first working day of the month, and all accounts which have statements produced on billing day 1 will always have them produced on that day, although the day of the month that is billing day 1 may be slightly different.

There are two broad types of data for each customer, those based on simple summary measures of the month's activity, and those based on new transactions made in the month. We call these 'account history' and 'transaction history' respectively. We list their elements below, but essentially the former is a fixed set of records captured or calculated (e.g. interest, or amount spent) each month. Every statement also records each transaction made in the billing month, and each of these individual entries is a fixed length record in itself, but any particular customer can make no transactions, or (in theory, at least) an infinite number.

3.1.1 Account history

The following information is contained in our data set.

1. Balance on the account, at the time the statement was produced.
2. The amount of last month's payment.
3. Interest applied to the account, if any. This is calculated based on a financial convention known as the 'average daily balance', which is derived as follows. Each day, the balance on the account is calculated, net of all new purchases, repayments, interest or other charges applied. These are then summed over the period since the previous statement, and divided by the number of days in the month to arrive at the mean daily balance.
4. Annual spending in the past 12 months.
5. Any status code that has been applied to the account. This is an indicator of

circumstances such as fraudulent use, repayment problems or account closure.

6. If there have been late repayments, how many days late the payment was made.
7. The credit limit allocated to the account.
8. The number of cards issued on the account. A card company has an agreement, with the 'main account holder' regulated under the Consumer Credit Act, that is legally binding on both sides. This card holder can request other cards to be issued, but any spending on those remains the responsibility of the main account holder. Typically this will be a married couple, one of whom will be responsible for payment, but they each carry their own individual card for use on the account. Another common scenario is where a parent requests an 'authorised user' card for a child, often for use in emergencies. Only one statement is produced (with occasional exceptions) regardless of the number of cards per account.
9. An indicator of whether the customer is a participant in Barclaycard's loyalty scheme.
10. The account number, if appropriate, of any other card accounts which are held with the company, as well as the Visa accounts which form the subject of this work.
11. Behaviour score. This is a derived variable, and is built up from a score card calculated from data on how each customer uses her/his account. It is expressed as a three digit number, but there is a monotonic transformation between score and the probability that a customer may default.

All of these fields are updated every month, but other information is available to us, most of which is not regularly updated, or at best, infrequently. It is as follows.

1. Date the account was opened.
2. Age of the main account holder. This is missing for around a quarter of Barclaycard's customers, because many have been with the company for more than 20 years, and there was no mechanism for capturing and recording age when they first acquired their cards.
3. Geodemographic indicator – the MOSAIC code, which is developed from the

Government's census data by Experian.

4. Sex of the main account holder.

3.1.2 Transaction history

1. Date the transaction was applied to the account. Note that this is not the same as the date the transaction was actually made, but usually a day or two later (it could be longer in the case of some foreign transactions). This is explained more fully below.
2. Amount of the transaction.
3. Merchant Category Code (MCC) of the outlet where the transaction was made. This is a four digit identifier, agreed by the members of the VISA International Service Association (the '*Association*' referred to in Chapter 2). In most of our work, we will aggregate these into a number of (what we will define as) *trade sectors,* which we describe more fully in Chapter 7. This is for two reasons, the first of which is that many MCC codes form natural groups. A good example of this is airlines, which have more than 300 individual MCC codes. Often, we are more interested in whether or not a card holder has bought an airline ticket, not which airline was the carrier. The second reason follows on naturally from this, because we believe that it will be easier to manipulate and analyse a relatively small number of approximately homogeneous sectors, rather than several hundred individual MCC codes.
4. Country where the transaction was made.
5. Transaction type. This is an indicator of such things as sales transactions, credit refunds, cash handling charges, type of cash transaction (over a bank counter or at a cash machine).

We expect that spending will be seasonal, and we further expect that the patterns will be different in different trade sectors but there is also a systemic distortion too. Every account (providing there is some activity on the account) has a statement produced each month. The account history is then updated, and all of the transactions recorded after the previous statement are added to the current one. This

means that many of the transactions made by customers whose statements are produced near the beginning of the (calendar) month will have occurred in the previous month.

It is for this reason that all of the spending data in this thesis have been treated differently from account history data. If we were to look at spending in statement months we would be introducing a distortion because of the way that statements are produced (as described in Section 3.1), which would result in a smoothing effect. For example, if we look at the uplift in spending at Christmas, data derived from statements shows that the peak occurs in January. We know, however, that much of the spending was actually carried out in December, but then recorded on statements produced in January. To avoid this complicating feature, we have grouped each customer's spending such that each analysis period covers the same dates, based on the date the transaction was applied to the account – known as the 'posting date' – not on the month the statement was produced. This period varies according the particular analysis, but is typically by day, by week, or some aggregate of these.

Also, all other things being equal, the greater the number of working days in a (calendar) month, the more transactions that Barclaycard will see in that month. The transformation just described has the benefit that it removes most of the distortion caused by varying month lengths.

There is a lag between the transaction date and the date on which that transaction is applied to a customer's statement. Some of the delay occurs because of the time it takes to process a transaction through the four party network described in Chapter 2. Most of the time this is not a problem for analysis, because the time delay is relatively consistent. A consequence of this, and the fact that transactions are (or

were, at the time these data were extracted) 'posted' to the account usually on working days, is that – apparently – little spending takes place at weekends, or on public holidays. This is not the case, of course, but we need to be aware of this in any analysis. It is known, for example, that most spending at a weekend is applied to statements on the following Monday. The main exceptions to this are as follows.

- Bank holidays. These can introduce a delay of four days (at Christmas or Easter), rather than the normal two, which is the case most weekends.
- Transactions that are made overseas. Some of these can take several weeks to reach the account.

Another distortion could occur if we seek to analyse spending by statement month if the last days of the month fall at a weekend. To illustrate, consider the weekend of the Bank Holiday on the 1st of May 2000. All of the spending that Barclaycard customers made (on their Barclaycard) on the Saturday and Sunday of that weekend (April 29th and 30th) will be recorded in the following month, because the first working day after this weekend is the 2nd of May. Transformation of monthly calendar dates into (say) four week periods still does not resolve this issue, although given the data limitations, there is little else we can do to improve things still further.

The final reason we make such adjustments is that the company allocates billing days to customers, not the other way round (although some people request that their billing day is changed). We expect to see different patterns at different times of the (calendar) month, which would coincide with weekends, or peoples' pay days, for example, and we may seek to describe or model such features. If we looked at statement months rather than calendar months, some of these transactions could apparently be in the 'wrong' month.

These are all complicating features in the data, and were referred to as 'data distortion' by Hand and Blunt (2001). The spreading of the billing days over the month makes the data mining task harder, although it illustrates the secondary nature of much data mining. If Barclaycard's operations had been designed with data mining as the primary aim, then all statements would be produced on the same day. This would cause massive difficulties in operating the business, given that several million statements are produced each month.

3.2 Other data issues

3.2.1 Customers 'lost' from the sample

Another problem is lost or stolen cards. Our sample was selected by taking accounts that had '1' as the 11^{th} and 12^{th} digits of their account numbers. The number that comprises these two digits is, and always has been, sequentially allocated to new accounts, from a value in the range 00 to 99. The sequence runs ... 98, 99, 00, 01, ... as 99 is approached and passed. This is a common method of sampling at Barclaycard, and every time such a sample has been taken, and then checked, it has found to be representative of the file as a whole. We can thus be reasonably confident our sample contains representative numbers of accounts by age of account, and the behaviour of those accounts.

Around 5% of customers are given a new account in a year because their card is lost or stolen. They then only have a 1 in 100 chance of reappearing in the sample. The converse is true, that the same proportion of such accounts appears in the sample in the course of a year. The upshot of this is that two samples, selected using the same two digits, but in subsequent years, will have 10% of customers which are different,

but nevertheless are existing accounts (5% leave, and another 5% appear). We might expect a further 10% of 'churn' because of attrition from the file, and new customers opening an account for the first time. In Section 3.1 we described the two data extracts we had taken, at different times, and this time difference meant that we 'lost' 398 accounts. Notice that, if the loss in any 12 months is around 10%, we would expect a 4% loss in five months (our second extract covered August to December), and 398 accounts is 3.9%, exactly in line with expectations. Here, we are not concerned with the concomitant gain in accounts, because we were not able to go back and select the previous history of those who 'appeared' in the second extract.

We also 'gained' some others, whose accounts were either opened, or who appeared in, our sample in the five months from August 1997 to December 1997. As we have no more data from these customers, we will not mention them again.

This effect can be mitigated by examining the dates on which each account was opened, and comparing it to the date on which the account appeared in (or disappeared from) our sample. We can then select those accounts for analysis that were present only for the whole of our analysis period. Problems can arise because customers who need their cards replacing tend to be more active than the average, and by leaving them out we will be ignoring some potentially useful data. If we include them, though, we risk ascribing their sudden appearance (or disappearance) to a change in behaviour. This is obviously not the case. In the following work we have omitted them, but it could be the subject of future work to examine these customers separately.

3.2.2 Selecting only those customers present for the whole of the period

Although this makes sense from the point of view of constructing models, where we would not want to record disappearance from our sample as a change in behaviour, there is a major consequence of this, as well as the statistical benefits. The people we choose to omit from our sample in the manner described in the last section are still customers, to whom Barclaycard provides a continuous service, apart from the day or two in which their card is being replaced. By omitting them, we implicitly assume that they are the same as those who stay in the sample. We have no reason to make this assumption, and other work done by Barclaycard does indeed indicate that customers lost because of our sampling procedure tend to be more active than the average.

3.2.3 New customers

A further group we need to consider is new customers. Such accounts usually have different usage patterns compared to established ones, and they are often analysed separately. However, in this work we were concerned mainly with established accounts, for two reasons. Firstly, much analytical resource is devoted to new accounts, especially bad debt in the early years, so a lot of work we could do might duplicate that of others in the company. More importantly, our aim was to look at established accounts and our sample had a relatively small number of new customers (793), and we would ideally select more than this for any work on this group. This is the number of new accounts we would expect in a sample of our size, given the proportion in Barclaycard's database at the time of our extract.

3.2.4 The sample that remained

After removing new, closed and 'lost or stolen' accounts from the sample, we were left with 7,944 people whose accounts had been open for the whole time, and were in the data set each month. In much of the analysis that follows, we have split this into two equal sized samples of 3,972 customers (we selected customers for each group by giving each one a random number, to the level of precision allowed on our computer, which was to 17 decimal places). We call one of these sets our 'design' set, on which we will attempt to devise predictions and classification rules (see Chapters 4-6). The other we call our 'test' set, which we will use to assess the rules we have devised.

For the prediction of spending behaviour (see Chapter 9), we needed to use data from August to December 1997, as well as from January 1996 to July 1997, so our samples shrunk a little, to 3,827 and 3,826 people in the design and test sets, respectively. The loss of 'established' accounts was thus 291 customers, less than the 398 who were lost in total (described in Section 3.2.1). There are always a number of people who open and close accounts quickly, and who will change from a new account to a closed account in less than a year.

3.2.5 Negative balances

Typically, most customers will repay some, or all, of the balance on their statement. There is no reason to pay more than the amount shown. In theory, this should mean that we never see a negative balance, but there are customers who exhibit such characteristics. For example, in one part of our sample of 3,972 customers (see Section 3.2.4), if we look at 12 months of statements, there were 594 occasions in the year where accounts had a negative balance (1.2%), among 188 customers (4.7%).

Within this, most customers had only one or two (117 of the 188) but there were 9 people who had a negative balance for the whole year.

Some of these are for small amounts, and are caused by what is known as 'small balance write offs', frequently an indication that there has been a problem with an account, and that it may be closed soon. Many, however, are not, and to illustrate with two extremes, why should a customer repay more than £3,000 when her balance was only £527? This seems easy to answer, because a purchase of £2,125 was made the next month, and so presumably the transaction took longer to reach the account than the customer expected, or she had the money for the purchase available and paid it into her account 'pre-emptively'.

Many of the large negative balances last for only a few months, and so this would seem to be a reasonable explanation. It is not possible to ascertain the real reason unless we interview customers about their behaviour when this happens. We could ask a sample of these customers, but such behaviour does not cost the business anything, so there are few financial benefits to finding out why a small number of customers behave in this way.

Another customer had an average balance of -£2,089 across the year, and the 'highest' value it reached was -£1,096. In several months he did not make a repayment, but did make more transactions. We are at a loss, though, to explain why, in the month when his balance rose to its highest, he spent £100, and then repaid £2,000! In this case, the balance is large and negative (i.e. on the 'wrong' side of zero balance), and is obviously an outlier. Obviously, if we do not treat these customers' activity with care, any modelling could be affected. In the classification work, one of the explanatory variables we use is the ratio of the balance to the credit

limit, and for customers such as the ones just described, we set the figure to zero, and although this is somewhat arbitrary, it is commonly used.

In Chapters 7-9 we describe some distant points in highly skewed distributions, which could also be thought of as outliers, but we believe that these are examples of the sparsity of the population in the tails of these distributions. In the latter cases, outlier identification is not a trivial task, and cases must be examined individually, according to the nature of the outlier.

3.2.6 Caveats

There is one other caveat that should be noted. We have data on one card only, not on competitor cards, and likewise, we have no information on alternative methods of payment such as cash or cheques. This means we know nothing about other products or services our card holders may use for their borrowing or spending activity: these could include personal loans, overdrafts, debit cards and cash.

Chapter 4

4 Repayment behaviour – descriptive models

4.1 Introduction

In this chapter we investigate the different patterns of repayment behaviour that we see in our data, and will examine, and describe, some relatively simple ways to summarise such patterns. We have two aims: to be able to find descriptions of behaviour, and to classify customers according to such behaviour, then to use this as the basis for incentivising suitable groups. There are undoubtedly many possibilities, but we will select the simplest and most obvious at this stage. If we can achieve a high level of predictive accuracy (in some sense yet to be defined) from any relationships we deduce, from any of these descriptions, we will be able to avoid using more complex ones. This matters because there are three important issues if we hope to implement any rules that we devise.

1. As few as possible predictor variables should be used, to minimise the time taken to collect and clean the data, which will enable the business to respond more quickly. See Chapter 3 for a description of the data and some distortions in it.
2. The models should be as simple as possible, so that they can be implemented by people who might not be statisticians.
3. The rules we devise will probably need to be programmed into mainframe systems. The simpler they are the better, to keep development time and cost to a minimum.

Our objective is to be able to determine how accurately we can predict behaviour, and to do so we use a variety of techniques. In this chapter and the next, we will

consider simple linear regression, classical linear discriminant analysis and *k*-nearest neighbour methods.

The initial regression work is to allow us to describe different groups, which we can then treat differently. If we can devise a suitable grouping scheme, the problem will be amenable to discriminant analysis. Some method of grouping can be quite important for implementation by credit marketers, because it will allow us to 'clean' our data set of anomalous observations, and simplify it greatly. Some examples of anomalies were described in Chapter 3, and we give more in Section 4.2.6.

We also make the important distinction between 'type 1' and 'type 2' problems, which we defined in Chapter 1. The former permit us to treat people differently, while the latter are concerned with behaviour modification (perhaps by a suitable incentive, for example).

4.1.1 Full and partial repayment

All credit cards of which we are aware have, in their 'Terms & Conditions', a statement such as 'you must make the minimum payment by the due date shown' or 'you must make either the minimum payment shown on the statement or any larger amount chosen by you'. All credit card issuers charge interest on balances from the previous month until a payment is made, and many levy a fee if a customer is late with a payment. If partial payment is made, the card holder will be charged interest on the balance left outstanding, and usually on all new transactions from the time they were made. On most cards, customers who repay in full do not incur interest, although there are one or two exceptions.

To a card issuer, the most profitable customer is one who makes the minimum repayment as late as possible in the billing cycle, and leaves the balance outstanding for the maximum possible amount of time. We believe that some of the newer entrants to the UK market have many more customers in this category than the 'traditional', longer established, issuers. This is certainly the case for those which came to this country from the USA (as we described in Chapter 2). One way to measure this is to consider the ratio of outstanding balances to spending – the higher this is, the greater is the proportion of total balances that earns interest (as distinct from 'full payers' spending, which earns the issuer little, or no, interest). For example, in the UK, this ratio is 0.37 (BBA, 2001), but for the largest 50 credit card issuers in the USA it was 0.54 in 2001 (The Nilson Report, Number 732, *www.nilsonreport.com*). Some issuers are even higher: Providian, for example, which also issues credit cards in the UK, had a ratio of 1.02 in America. However, that in itself caused problems for Providian, which replaced its Chief Executive in October 2001, and was widely rumoured to be for sale.

In any given month, around half of Barclaycard's customers repay their full balance. Every time this happens, the company earns no interest on the amount that a customer spent in the previous month, but has to fund the balance until the customer's payment is received. To an issuer, it is therefore in its interest to encourage fewer customers to pay in full, and more to borrow for as long as possible (without encouraging people into debt which they cannot afford, however). In this thesis, we will not consider risk or 'bad debt', although this has been the subject of much work over many years (for example, see the 'Statistics in Finance' internet site

of the Consumer Credit Research Group at Imperial College, London at http://stats.ma.ic.ac.uk./creditgroup/financelist.html).

In the next section we describe some techniques that we can use to summarise the behaviour of different types of consumer, and how they make repayments to their credit card accounts. We will then classify these customers, and attempt to model behaviour which might be of benefit to a card issuer. The intention of this is to try to achieve a balance of predictive accuracy and simplicity, seeking techniques which can be expressed as simple algorithms, as we discussed in Section 4.1. In much of the work in this chapter and the next, we have split our sample into two equal parts, with the aim of predicting behaviour in the second half from that in the first. These parts were of nine months each, and were February to October 1996, and November 1996 to July 1997. The reason for starting in February is simple: interest is applied to a statement for behaviour that the card holder exhibited in the previous month, as described in Chapter 3. A customer makes a repayment in the current month against the balance as it was on the previous month's statement. For data based on customers' billing months, this will always be a problem – we will not be able to use much of the first month of data, because some of the key measures are based on activity that took place before the data were extracted.

4.2 Selecting a suitable classification scheme

4.2.1 Use of linear regression to predict the amount of interest paid

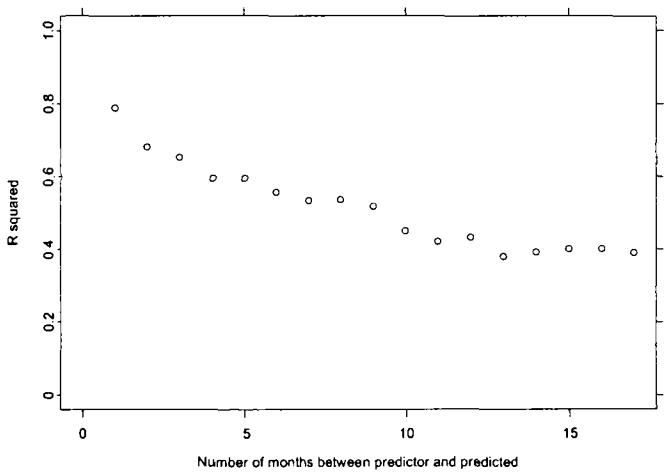
We start by modelling interest because, for any credit card company, interest is one of the major sources of income, so the ability to be able to predict its amount will be valuable. We described, in Chapter 3, how interest is calculated and added to a

customer's monthly statement. Let $t = 1$ be the first month of the sample, $t = 2$ be the second, with $t = 18$ being the last, and let x_t be the interest applied to each customer's statement in month t, and perform the simple linear regressions as follows.

$$\mathbf{x}_{18} = \alpha_{0t} + \alpha_{1t}\mathbf{x}_{18-t} + \varepsilon \quad \text{for } t = 1, \ldots, 17$$

We show the R^2 that resulted from these 17 regressions in Figure 4.1, where the outcome vector in each case is the interest applied in the final month of our period. It shows quite clearly that the R^2 is larger the closer is the predictor month to the final month. However, when twelve months has elapsed between predictor and outcome, there is little further deterioration, suggesting that there may be a group of customers who are regular interest payers (we investigate this more fully in Section 4.2.4). The better fit in contiguous months is to be expected, because we would expect people's behaviour to change less over the short term than over a longer period.

Figure 4.1 R^2 and its relation to time between predictor and predicted interest



Choosing other combinations of interest paid confirmed our supposition – that we can achieve reasonable predictions of interest by using earlier values of the same

variable. We also modelled the total amount of interest in the second nine months from the first nine months; and used a matrix of individual months of interest in the first half to predict the total amount in the second nine months.

The closest month of interest, in a temporal sense, is virtually all that is necessary to achieve the best fit, and other variables (including the other account history variables) add little extra predictive power. We would expect this to be the case, however. Customers who have a large balance may not be able to repay much each a month, but only a relatively small proportion. For such customers, there should be a close relationship between interest applied in successive months, because the balances from one to the next are likely to be similar. This has implications for the business, because it implies that speed of response for producing incentives is likely to be important.

While this approach will allow us to identify who pays a lot, or a little, interest (or none at all), it will not allow us to identify how often people pay less than the full amount of their balance.

4.2.2 Number of occasions on which interest is incurred

We are also interested in the number of *occasions* on which people make a partial repayment, and will thus incur interest, and predicting this is a 'type 2' problem, as we defined it in Chapter 1. However, one aim of the business might be to predict those who might be incentivised to make the amount of interest paid a little higher on one (or more) of those occasions. This is a slightly different problem, because we first need to identify the customers who pay interest, and then to identify those who might be more amenable to paying a little more. Both ways of incentivising

customers are important to the business, as both would lead to extra revenue. In the next two sections we describe a simple classification scheme, based on the number of times people incurred interest, and some of the advantages of using this simple summary number, rather than the amount of interest paid. The former is restricted to an integer between 1 and 18, the number of months in our sample, but the amount of interest someone may pay in a month is limited only by their balance.

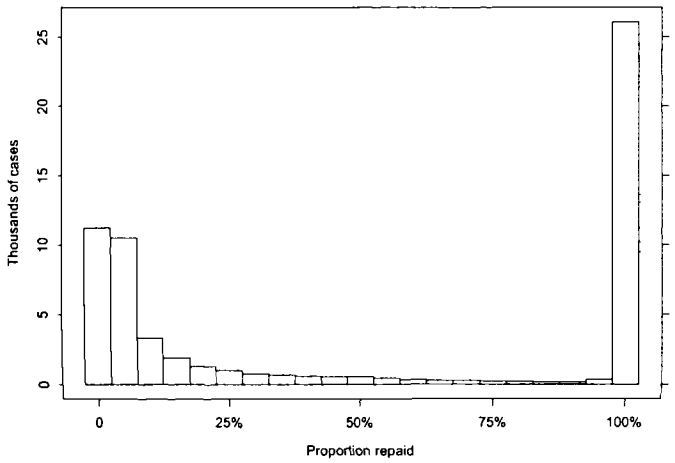
4.2.3 Devising a suitable classification schema

The simplest way to find the number of occasions on which people make a partial repayment is to look at each individual's payment in the current month and balance the previous month. If the payment was less than the balance, then interest will be applied to the account.

In Figure 4.2 we show the proportion of the balance repaid for all possible occasions and customers in our sample, but with zero and negative balances removed (if there was a negative repayment we set the value to zero). A negative repayment signifies that a payment was deducted from the account for some reason, usually where the repayment cheque has 'bounced' because of insufficient funds in the bank account from which it was drawn. Negative balances usually occur when the customer repays an amount which is more than the amount shown on their statement, a phenomenon we discussed in Chapter 3. The bars are at 5% intervals, but note that in the rightmost one, the interval is drawn as follows.

$$\text{Proportion repaid} \geq 100\%$$

Figure 4.2 Proportion of the balance repaid (bars are at 5% intervals)



The distribution is markedly bimodal, and most of the payments in the right hand mode are 100% of the balance, with a relatively small number being more than this. This pattern of behaviour is similar in every month, so 'full repayment or not' would seem to be a suitable place to partition our data for classification into full or partial repayment classes each month. There is a good business reason for trying to develop such a classification. It is that we wish to try to incentivise customers to make one (or more) extra partial repayment(s) in any particular period. The amount of that extra interest is of lesser concern (at least as a first stage) than the fact that we can generate some extra revenue from particular groups of customers. Any extra interest payments are incremental for the business. As a second stage, once we have successfully generated some extra interest payments, we might seek to increase the value of each one. Also, every time a customer makes a repayment which is less than the full balance, interest will be charged. It does not matter how much less than the balance, only that the repayment is not for the full amount.

Note that there is a small number of anomalous occasions for which we needed to make adjustments, and we show these in Appendix 4.1.

4.2.4 Patterns of partial repayment

As we described in Section 4.2.1, we (effectively) have 18 months of behaviour that we can use for modelling. Let **X** be a 3,972 × 18 element matrix (there are 3,972 customers in our design set and 18 months of data), where, each customer x_i is allocated the value $x_{t,i}$ at time t. The other elements of the definition are as follows: 'payment$_{t,i}$' is the amount of the payment made by customer i in month t; 'balance$_{t-1,i}$' is the amount of the balance on customer i's account the previous month, which we denote by t - 1.

$$x_{t,i} = \begin{cases} 0 \text{ if balance}_{t-1,i} \leq 0 \\ 0 \text{ if } \left(\text{payment}_{t,i} \div \text{balance}_{t-1,i}\right) \geq 1 \\ 1 \text{ otherwise} \end{cases}$$

The steps of this definition need to be applied in order from top to bottom, otherwise people could appear in classes that we do not intend. Then, for each customer i, let the total number of partial repayments each makes in the first and second halves of our sample, be n_{i1} and n_{i2}, as follows (the subscripts 1 and 2 denote the first and second halves of the sample, respectively).

$$n_{i1} = \sum_{t=1}^{9} x_{t,i}\,, \quad n_{i2} = \sum_{t=10}^{18} x_{t,i}$$

For instance, as we described in Chapter 3, there are a small number of people who made payments when their balance was less than zero (i.e. Barclaycard owed them money). This means that these customers are, at least for these repayments, neither

full nor partial repayers, but we need to allocate them to a class. For these purposes, we choose to allocate them to the 'full repayer' class in the month, although we could create one or more 'other' classes if we chose. An alternative approach would be to have a class such as 'no or negative balance', but this leads to another set of problems. Around a third of customers have a balance less than, or equal to, zero at some point in a year, and if we allocate them to a 'no balance' class in a particular month, we must decide how to calculate n_{i1} and n_{i2} for each of them. By adopting our approach each time this happens, that month will add zero to n_{i1} and n_{i2}. Our objective is to predict those who may be incentivised to make another partial repayment, and thus pay interest. If there is no balance, the customer will not pay interest, and so, as a consequence of our definition, 0 again represents 'no interest chargeable'. This definition is flawed, of course, because it includes the temporarily inactive with the full payers, but is consistent with our approach of trying to identify the partial payers. Further work should be undertaken to investigate the first sub-group of people.

Note that this broader definition of class 1 (each month) includes, by default, both of the following groups.

1. Customers who make a payment, but which is less than 100% of their balance. We will call these 'partial repayers'.
2. Customers who are late with payments, and so effectively make no payment at all in month t, but a double payment in month $t+1$. We will call these 'late payers'.

For these two more precise definitions, each customer x_i is allocated the following value at time t. As before, each row of this definition should be taken in sequence from top to bottom, and all of the other terms are as before.

Late payers:

$$x_{t,i} = \begin{cases} 0 \text{ if balance}_{t-1,i} \leq 0 \\ 1 \text{ if payment}_{t,i} = 0 \\ 1 \text{ if } (\text{payment}_{t,i} \div \text{balance}_{t-1,i}) \text{ AND interest}_{t,i} > 0 \\ 0 \text{ otherwise} \end{cases}$$

'Real' partial repayers

$$x_{t,i} = \begin{cases} 0 \text{ if balance}_{t-1,i} \leq 0 \\ 1 \text{ if } (0 < (\text{payment}_{t,i} \div \text{balance}_{t-1,i}) < 1) \\ 0 \text{ otherwise} \end{cases}$$

Thus there are three ways we can classify those customers who make less than a full payment – 'real' partial repayers, late payers, and the broader definition which includes both. We will investigate all three ways of classifying customers (the latter two in Chapter 5), because all are potentially interesting, and could generate extra income for the business. First, in Table 4.1, we show how customers are distributed according to the broader definition, rather than either of the more precise definitions. The rows show the number of ones in the first half of the sample period, the columns show the number in the second half. The correlation coefficient between the number of ones in the first and second halves, effectively the marginal distributions of the figures shown in Table 4.1, is 0.87.

We believe that the people who already make one or two partial repayments might be the best groups to target. They are the two biggest groups, apart from the extremes, and might be amenable to incentives, given that they already have slight divergences from wholly consistent behaviour. This latter point is important, as our objective is to encourage certain groups of customers to make an extra interest payment. The group with eight partial repayments is also large, but as most of them 'migrate' into the group with nine, they already seem to be borrowers in most months anyway.

There are two main ideas here: (i) identifying those who are not entirely resistant to paying interest, and (ii) identifying the subgroup of those who are already so totally relaxed with, or accustomed to, regular interest payments that incentivisation cannot make them pay more interest.

Table 4.1 Number of customers by the number of '1's in each 9 month period

First half	Second half 0	1	2	3	4	5	6	7	8	9	Total
0	1,043	180	60	21	10	8	6	12	6	0	1,346
1	169	74	36	28	8	7	7	7	5	7	348
2	62	32	15	21	12	7	4	3	4	18	178
3	34	17	10	13	3	3	7	5	9	18	119
4	16	14	9	11	7	8	11	10	15	20	121
5	14	8	12	10	8	4	10	8	8	27	109
6	8	6	6	10	7	5	6	9	11	30	98
7	8	3	6	10	3	8	4	11	13	59	125
8	7	3	5	5	7	11	18	18	25	89	188
9	7	8	13	12	23	23	43	54	89	1,068	1,340
Total	1,368	345	172	141	88	84	116	137	185	1,336	3,972

There are two striking things about this table, and the first is based on the marginals. Take the right hand column (the first nine months) as an illustration: it is markedly bimodal. There is a large peak at 0 and a large peak at 9. This suggests that there are two kinds of people: those who pay off their bill each month, and those who never pay it off in full. These two 'ideal' kinds are smoothed out because nobody is perfectly consistent in their behaviour, yielding people who try to pay their balance

in full each month, but occasionally do not, and those who essentially use the card for borrowing. The two types have about equal numbers of people.

The second striking thing arises from the bivariate distribution. The two marginals are essentially the same, but there are many off-diagonal elements (1,706, or 43% of the total), showing that individuals do not stay in any particular group. Indeed, there is no reason why they should, except for groups 0 or 9. That is, why should someone who paid 4/9 times in full in the first 9 months do the same in the second 9 months? This means that there is stochastic equilibrium, but individuals move around. Even in groups 0 and 9 quite a lot of people move around (about 300, or 25%, of those in these two groups in the first period move in the second). We might be interested in any behaviour migrations: for example, people who have 0 in the first nine months but then go to 3 or more in the second. More generally, we are interested in anyone who is consistently on 0 but then makes more than 1 or 2 partial repayments in quick succession, because we might speculate that they may no longer be able to service their debts.

Assume the number of partial repayments, x, made over the periods of our sample, be independent and follow a Poisson distribution with a constant mean. Further, suppose the mean number of partial repayments of individuals over the same periods differ according to a Beta distribution across the population. Thus, the frequency distribution of partial repayments should follow a beta-binomial distribution. Conditional on n, the number of repayments, this leads to the following.

$$P(x \mid n) = \binom{n}{x} \frac{B(v_n + x, w_n + n - x)}{B(v_n, w_n)}$$

Solving this equation for each row of the data (i.e. for each n) shown in Table 4.1 gives a transition matrix in Table 4.2, where each figure is the predicted proportion of people moving from each number of partial repayments in the first time period to the second. Figure 4.3 shows the predicted values for 0 and 9 partial repayments in the first period.

Table 4.2 Transition matrix for the beta-binomial model

Number of partial repayments

First half	Second half 0	1	2	3	4	5	6	7	8	9
0	0.78	0.11	0.05	0.03	0.02	0.01	0.01	0	0	0
1	0.50	0.16	0.1	0.07	0.05	0.04	0.03	0.02	0.01	0.01
2	0.37	0.13	0.09	0.07	0.06	0.06	0.05	0.05	0.05	0.06
3	0.30	0.11	0.08	0.06	0.06	0.06	0.06	0.06	0.08	0.14
4	0.14	0.1	0.08	0.08	0.08	0.08	0.08	0.09	0.11	0.17
5	0.14	0.09	0.07	0.07	0.07	0.07	0.07	0.09	0.11	0.23
6	0.09	0.07	0.06	0.06	0.06	0.07	0.08	0.09	0.13	0.30
7	0.06	0.05	0.04	0.04	0.04	0.05	0.06	0.07	0.11	0.47
8	0.02	0.03	0.03	0.04	0.05	0.05	0.07	0.09	0.14	0.48
9	0	0.01	0.01	0.01	0.01	0.02	0.03	0.04	0.07	0.80

Basic stochastic process theory enables us to estimate the proportion of people that we expect to find in each state (0-9) in the long run. Assuming a stationary first order Markov Chain the limiting behaviour, or stationary distribution, gives the long term fraction of time that the transition matrix spends in each state. It follows that this matrix of long term probabilities, π, will have the property $\pi = \mathbf{P}\pi$, where $\mathbf{P}$ is the calculated transition matrix shown in Table 4.2. Thus to find the limiting behaviour we need to solve, for each row i and column j,

$$\pi_j = \sum_{i=0}^{9} \pi_i p_{ij} \quad \text{for all } j$$

Subject to $\pi_j \geq 0$ and $\sum \pi_j = 1$.

Solving this we find:

$$\pi_j = (0.347, 0.070, 0.043, 0.033, 0.029, 0.027, 0.029, 0.034, 0.052, 0.335)$$

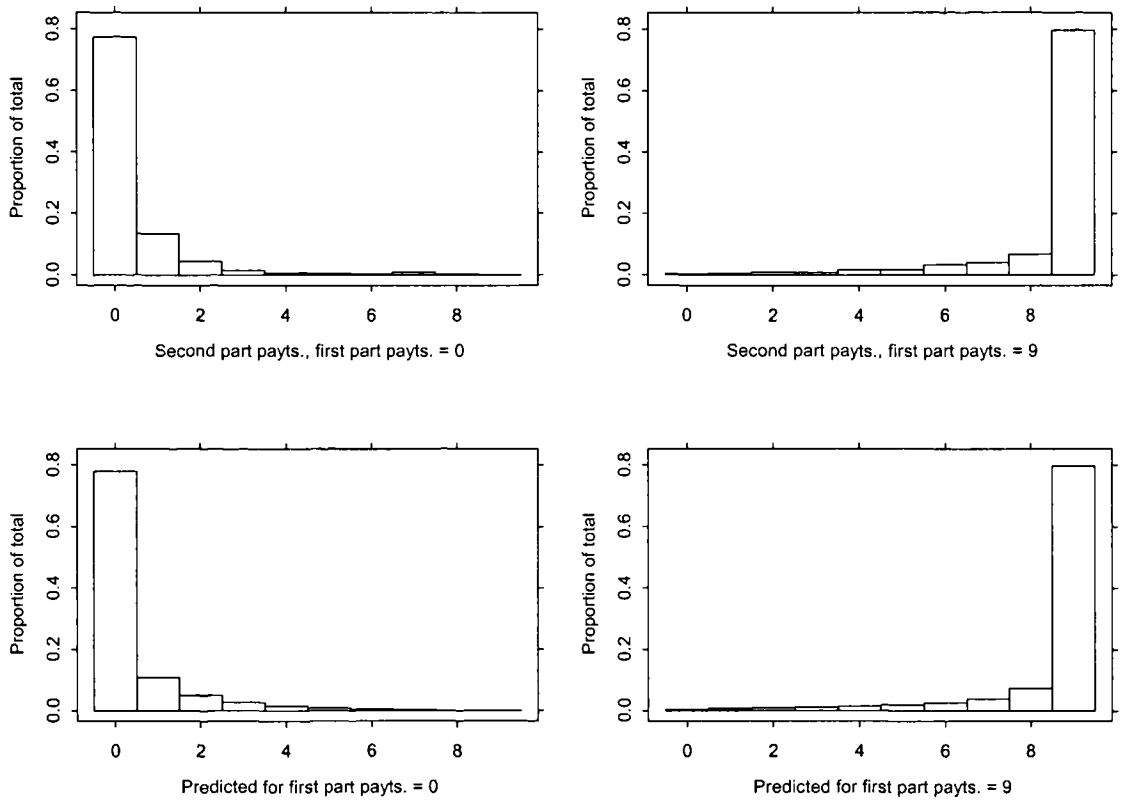
for all $i = 0, 1, \ldots, 9$ partial repayments.

Note that there is a crucial point for the business in these values. It is that, in the long term, around a third of customers will not make a partial repayment, nor a late one. The former is an important source of income for a card issuer, and the implication is that the company must be able to offer other products and services to this group if they are to generate much revenue. We described *interchange* in Chapter 2, which is designed to offset the cost of full payer 'borrowing' over the short term. This is fixed at just over 1%, but the cost of full payers' spending at a (London Inter Bank Borrowing Rate – usually abbreviated to LIBOR) rate of circa 5% is likely to be of the order of 0.5%. The difference between these is a narrow margin to generate income, once processing costs are taken into account.

In Figure 4.3, we show the observed and predicted values, for two values of partial repayments in the second half of our sample, 0 and 9. By eye, the fit appears to be very good, but the χ^2 statistic for this model is 117.8, and the probability of obtaining a value this large from a χ^2 distribution with 56 degrees of freedom is very small indeed. Thus, the observed and predicted values are significantly different. This is the first instance that we will see, of several throughout this work, where the fit predicted by a model is significantly different from the observed values. Examining

the fit visually would lead us to conclude that the observed and predicted values are very close, and in fact, the model is perfectly adequate for our purposes.

Figure 4.3 Observed and predicted for partial repayments = 0 and 9



The choice of 9 partial repayments in the above analysis had no special significance – we adopted it for convenience, given the data we had. Using the results from the previous paragraph, and assuming that the limiting behaviour is an adequate model for the proportion of each 9 months that customers make partial payments, we have a distribution showing long term the proportion of the population who make {partial repayments none of the time, 0.111 of the time, 0.222 of the time, ..., all of the time}. By interpolation, we can now deduce the continuum of values between 0 and 1, which could prove to be immensely valuable. We cannot say things about

individuals with great accuracy, but we can compute the long term number of times that n customers pay interest. Although the predicted and observed are significantly different, the predicted values are close enough to be usable in business context and we will describe this more fully in Chapter 5, where we use these results successfully to develop a classification rule for an unseen test set of data.

We performed the previous analysis using the number of partial repayments, 0, ..., 9 in two successive nine month periods, but we could aggregate customers into two groups: those of interest (i.e. with one or two partial repayments) and all others. However, this would be equivalent to saying there are two kinds of people, with different probabilities of not making a full repayment. In reality, of course, we believe that each customer has a slightly different probability p of skipping a full repayment, so the 10 by 10 matrix is a better representation because it captures the fact that there are different kinds of people.

Let p_i be the probability that a customer makes i partial repayments. We are interested in trying to identify those who might respond positively to incentivisation schemes. What such an incentivisation scheme does is to try to shift the stationary distribution towards higher values, which we seek to achieve by increasing the values of p_i. We acknowledge that it might be difficult to increase p_0 so we focus attention on the p_i with $i = 1$ to 8. Also, it might be difficult to move those customers at the higher values of i because they are already close to always making a partial repayment, and may be unlikely to make one more.

4.2.5 Patterns of partial repayment – 'real' partial repayers

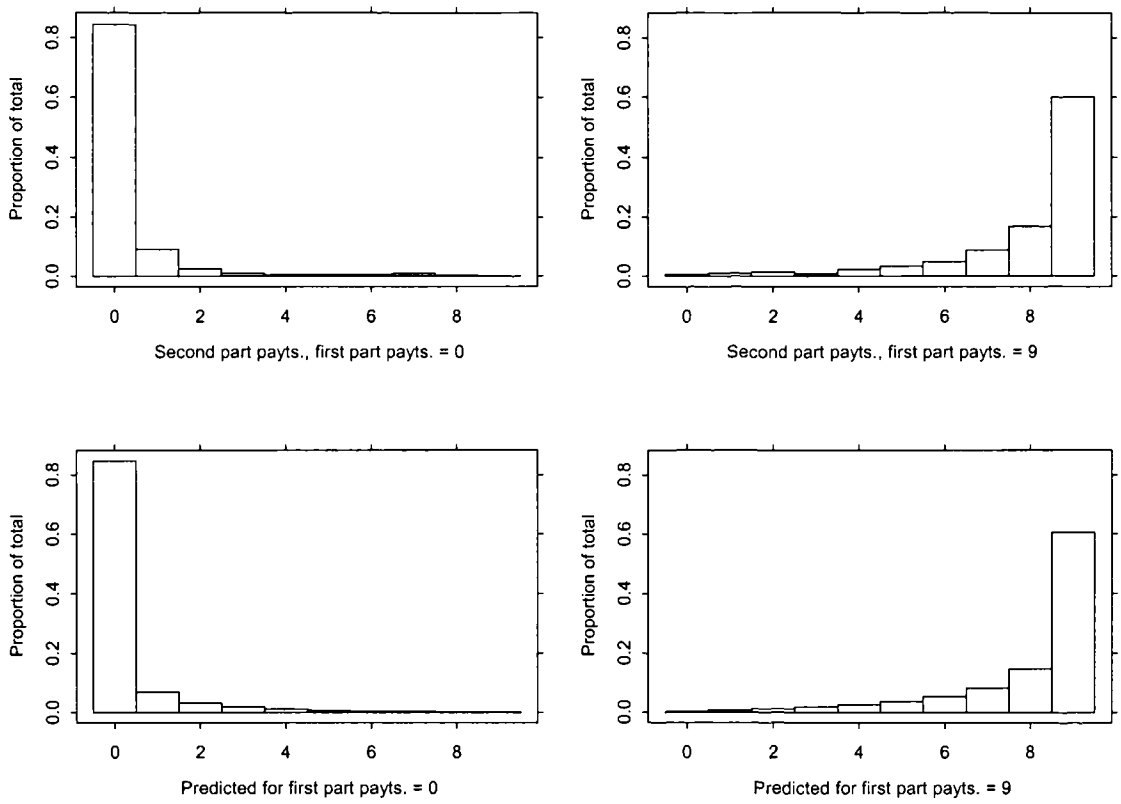
In Section 4.2.4 we described three ways of classifying customers who make partial repayments, and most of our work used the 'broadest' definition, which was to classify any instance of less than full repayment as partial. It included those who made a partial repayment and those who were simply late, and missed the payment due date (see Chapter 2 for a description of this). In this section we demonstrate the predictive power that is possible with a more precise definition – that of 'real' partial repayers. These customers make less than a full repayment in a month, and exclude those who are simply late with a payment. We show their number of partial repayments in the first and second halves of our period in Table 4.3.

Table 4.3 Number of partial repayments in each 9 month period

First half	Second half 0	1	2	3	4	5	6	7	8	9	Total
0	1,341	144	41	16	8	9	9	14	5	-	1,587
1	153	40	17	18	9	8	8	7	6	6	272
2	52	15	15	14	6	4	8	7	2	14	137
3	25	13	12	11	9	8	11	19	9	8	125
4	19	9	12	12	22	16	12	15	14	8	139
5	19	12	4	16	17	29	26	21	18	23	185
6	9	6	6	12	18	33	37	41	33	28	223
7	11	6	8	13	12	41	46	59	53	54	303
8	5	3	5	9	12	23	38	50	82	85	312
9	5	7	10	6	16	23	34	60	115	413	689
Total	1,639	255	130	127	129	194	229	293	337	639	3,972

The observed and predicted values are shown in Figure 4.4, and again we can see the same good visual fit. As before, however, the predicted values are significantly different from the observed values.

Figure 4.4 Observed and predicted values for 'real' partial payments = 0 and 9



Following the same steps that we did in Section 4.2.4, we can model the data using a beta-binomial distribution and calculate a transition matrix. Assuming a stationary first order Markov Chain the limiting behaviour, or stationary distribution, gives the long term fraction of time that the transition matrix spends in each state. It follows that this matrix of long term probabilities, π, will have the property $\pi = \mathbf{P}\pi$. For these data, this results in the following.

$$\pi_j = (0.451, 0.062, 0.044, 0.040, 0.041, 0.044, 0.049, 0.057, 0.071, 0.141)$$

for 0, 1, ..., 9 partial repayments.

4.2.6 Patterns of partial repayment – late payers

Repeating the approach for our third category of 'intermittent interest payers' we see the late payers' data in Table 4.4. Note that this is rather different from the previous two, with almost half the customers in the (0, 0) cell, which we would expect. We described the 'payment due date' in Chapter 2, and each time a payment is received after this date, a customer is breaking the terms of the agreement they signed when they took out the card. Most customers do not do this, because, at best, they will be charged interest on the outstanding balance on the statement, and they could be charged a 'late payment' fee as well.

Table 4.4 Number of forgetful payers

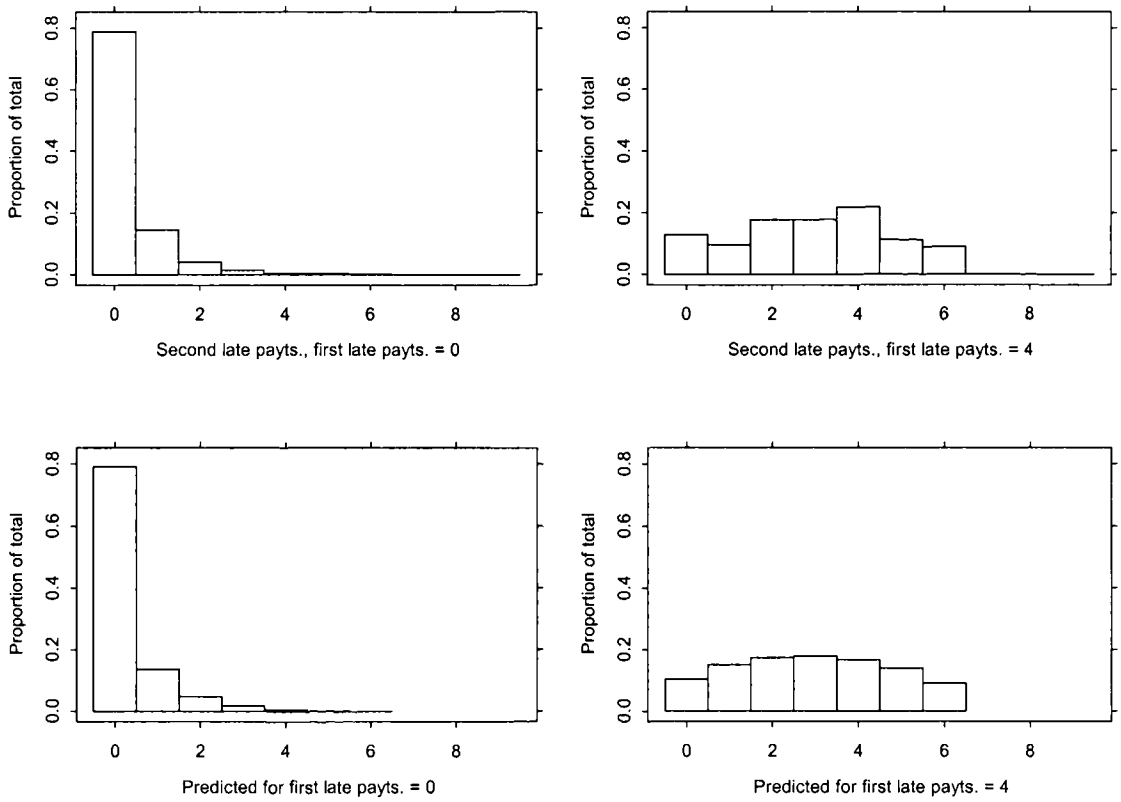
First half	Second half 0	1	2	3	4	5	6	7	8	9	Total
0	1,898	351	98	35	11	8	3	-	1	-	2,405
1	319	214	104	46	22	4	2	-	-	-	711
2	98	109	95	64	33	14	3	-	-	1	417
3	31	37	66	40	37	11	8	1	-	-	231
4	16	12	22	22	27	14	5	4	2	-	124
5	5	3	12	12	10	17	4	-	-	-	63
6	3	0	3	1	6	2	2	1	-	-	18
7	-	1	-	1	1	-	-	-	-	-	3
8	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-
Total	2,370	727	400	221	147	70	27	6	3	1	3,972

Again following the same steps as in Sections 4.2.4 and 4.2.5, we can model the data using a beta-binomial distribution and calculate a transition matrix. Assuming a stationary first order Markov Chain the limiting behaviour, or stationary distribution, gives the long term fraction of time that the transition matrix spends in each state. It follows that this matrix of long term probabilities, π, will have the property $\pi = \mathbf{P}\pi$. For these data, this results in the following.

$$\pi_j = (0.578, 0.182, 0.105, 0.065, 0.039, 0.022, 0.010, 0, 0, 0)$$

for 0, 1, ..., 9 late repayments.

Figure 4.5 Observed and predicted for late payments = 0 and 4



4.3 Conclusions

We have seen that ways do seem to exist of classifying people according to their repayment behaviour. With the objective of the business in mind, in Chapter 5, we will seek to predict, at a later time, those customers who make a partial repayment. We have developed a powerful tool for the business to use: the fact that we can predict – in the long term – the proportion of customers who make partial repayments {0, 0.111, 0.222, ..., 1} of the time.

We have been able to divide the types of repayment into three categories, again all useful for the business, and have made three important discoveries. The first was from the broader definition of partial repayers, and the most important cell in this matrix was the one at (0, 0). The implications of this are important for two large income streams – interest and 'out of order' fees, such as those imposed on customers who are late with payments. We found that more than a third of customers (0.347) will never pay interest or make a payment late, so any income the business derives from these customers will have to come from other sources. This could be interchange (we described this in Chapter 2) or from the sale of other services (Barclaycard travel insurance, for instance). As well as this, we also learned that a third (0.335) of customers will *always* earn the company some interest or fee income.

The second was among those customers who make less than full repayments, and we learned that one in seven customers (0.141) will always make a partial repayment, and almost half (0.451) will not, but that others do from time to time. We investigate the possibility of predicting (again, at a later time) these classes in Chapter 5, and then being able to incentivise such customers.

Finally, we found that the late repayers had, as we suspected, different characteristics from those of the partial repayers. However, a number of these customers may pay fees regularly because, once or twice a year, they are late with their payments.

Appendix 4.1 Anomalous values for which we made adjustments

In any particular month, some accounts have negative or zero balances, and a few have negative payments. The latter happens when a repayment cheque bounces. We made the following arbitrary, but hopefully sensible, allocations.

1. If there was a negative or zero balance, the customer was allocated to class 0. Negative balances happen when, at some time in the past, a customer paid an amount that was greater than the balance.
2. An issue that will have to be addressed in the future is what to do with accounts such as these, from which we expect no payment, but the customer is a regular user of his or her card. Having said this, there was one customer who maintained a negative balance for two years by consistently making repayments that were larger than his purchases.
3. If the repayment was greater than the balance, the customer was allocated to class 0.
4. If the payment was negative, the customer was allocated to class 1. This would usually happen when the customer's payment cheque has been returned unpaid, and so she or he has effectively not made a repayment.

Chapter 5

5 Repayment behaviour – predictive models

5.1 Introduction

We outline three different kinds of strategies for classification. Using the ten classes of partial repayers described in Chapter 4 (for the three different definitions of partial repayment we showed there), we seek to identify customers in the subset we believe will be amenable to incentivisation. We take two broad approaches to the problem. The first is to predict membership of each of these ten classes and then split the predictions into two: those customers whom we seek to incentivise, and all others. The second approach is to group the classes a priori into the two groups of interest, and then predict membership of these two classes. For each of these approaches, we explore more than one predictive model – linear regression, classical linear discriminant analysis and k-nearest neighbour methods. For the first of these, we will use the R^2 statistic as a way of comparing different regression models, while for the discriminant analyses we will also discuss suitable ways of assessing the models developed.

If the classes in a discrimination problem are defined by partitioning an underlying continuum (or, more generally, underlying continua), then there are two possible strategies for prediction. We could predict this continuum (using regression analysis or some modern variant) and then partition the predicted values, or we could assign the (design set) points to classes first, and then seek simply to predict class membership (or, more correctly, to predict probability of class membership at each pattern of covariate vectors), using some supervised classification approach (such as

linear discriminant analysis or logistic regression). The first strategy makes use of more information - the actual numerical values of the response variable - but can only do this at a cost of making some assumptions about the conditional distribution of the response variable given the covariate values. If these additional assumptions are wrong, then the predictions are likely to be biased. The second strategy (predicting the already partitioned response) is less powerful, but makes fewer assumptions. Hand, Oliver, and Lunn (1998) describe these ideas, but not the best methods for actually carrying out these sorts of analysis.

5.1.1 Classification of repayment behaviour

In the last chapter, we described different patterns of repayment behaviour that we see. We also speculated that the most likely groups of people who would respond to certain kinds of incentive would probably be those who already make one or two partial repayments. We therefore seek to develop classification rules so that we can allocate people to the correct groups at a time in the future.

5.1.2 Using a simple assessment table to judge our predictions

We have three definitions of partial repayment – late payers, 'real' partial repayers, and a broader one that encompasses both of these. All are valuable from the business' point of view, because they allow us to classify customers according to the income they generate, and so help earn revenue for a credit card issuer. Each of these definitions is summarised by a single number, between zero and nine, which is the number of partial (or zero) repayments made by each individual. Our objective is to predict peoples' membership of these 10 groups, for each of the three definitions. We discussed, in Chapter 4, that the behaviour of some groups of customers will be difficult to shift, so we will concentrate on people who make one or two partial

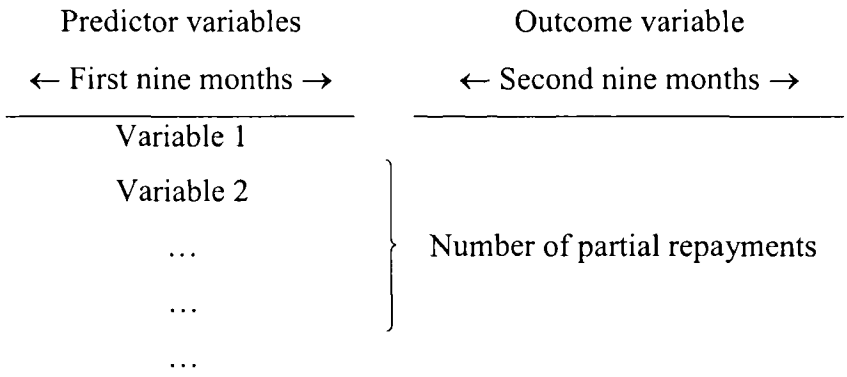
repayments. The business can take advantage of the behaviour of this group of people, if, by offering suitable incentives, we can encourage them to make another interest payment across the course of a year. We further believe (supported by the data) that we will not be able to change the behaviour of people who never make a partial repayment; and similarly, those who never make a full repayment cannot make any more partial repayments.

To enable easy comparison of different analyses, we will use Table 5.1, with the customer in class '1' if she or he makes one or two partial (or zero repayments), and in class '0' otherwise. Note that the tables summarise the results of our modelling, where the aim is to predict membership of the ten groups of partial repayments, 0, ..., 9. This means, of course, that in our 2×2 summary tables, class 0 contains 'hard core' full payers, 'hard core' borrowers, and customers in intermediate positions too. However, this heterogeneity is unimportant here, because we are only interested in those customers who made one or two partial repayments for this exercise, because we believe that these customers are more likely to respond to incentives than other customers.

In Chapter 3 we described how we divided our sample of 7,944 customers into equal sized design and test sets, and that we have eighteen months of data available. Further, for the classification work in this chapter, we have split those eighteen months into successive periods of nine months each. In the work that follows, we adopt a slightly more stringent approach than might be usual in a classification problem. First, we use data in the design set in the *first* period to predict repayment behaviour in the *second* period. None of the variables we use to develop the predictive rule span the second period, as illustrated in Figure 5.1. This might seem

an obvious thing to do, but many studies predict on a holdout sample of the data. Their models fail to take into account any variation arising from change over time, and thus tend to give over-optimistic predictions.

Figure 5.1 Construction of predictive rules



Similarly, when we seek to predict behaviour in our test set, we always perform this prediction for behaviour in the second half of the period, based on data for the first half, using the predictive rule developed using only design set data from the first half.

Table 5.1 Assessment of predictive rules

		True		
		0	1	Total
Predicted	0	*a*	*b*	
	1	*c*	*d*	
	Total	*C*	*D*	

We will then compare the two ratios $d/(c + d)$ and D/(C + D), and this comparison will allow us to form an assessment of the efficiency of the rule for targeting purposes. The larger the difference between the two ratios, the more efficient it will be when using the rule for the business. We also need the sum $(c + d)$ to give us a

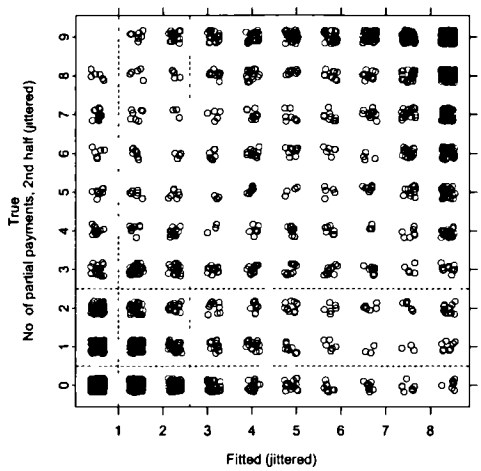
large enough group to use for targeting. We will demonstrate this in the discussion that follows the results.

A more common measure of misclassification might be the odds ratio (ad/bc), although there are many others too, such as error rate: see, for example McLachlan (1992), or Efron and Tibshirani (1993) for many references on these. In our particular case, we want to measure the ability of a rule to classify correctly membership of the predicted class, compared to the customer file as a whole. The odds ratio is a single degree of freedom summary, and is a comparison between the predicted classes. In many cases, the higher the odds ratio, the higher the $d/(c+d)$ ratio, but this is not always the case. Here, three measures are of interest: d, $d/c+d$, and the relationship between $d/(c+d)$ and $D/(C+D)$, and using the odds ratio gives us no information about the first two.

5.1.3 Linear regression, broader definition of partial repayment

Let $y = ax + b + \varepsilon$, where x is the number of partial repayments in the first half of our sample, and y is the number in the second half of the sample. First, we use the broader definition from Chapter 4. We find that $\hat{y} = 0.87x + 0.53$, with $R^2 = 0.76$. Our objective is to predict class membership, so choosing where to put the cut point to yield these predictions is part of the problem we face, but the R^2 is independent of this. The plot of the fitted values is shown in Figure 5.2. We have jittered the points to illustrate the density at different points on the plot, because, left unaltered, the underlying pattern is unclear because many of the points overlap. The position of the cut points (in this case) is obvious, and we used the vertical lines shown in Figure 5.2.

Figure 5.2 Linear regression, only one predictor variable



Summarising these results in our classification table, for both design and test sets, gives the results in Table 5.2. There is similar performance in the design and test sets. This means that linear regression is robust at this level of predictive accuracy, because there is no degradation in performance when the 'rule' is applied to the test set of data. However, this is mainly because the model is simple and the data set is large.

Table 5.2 Linear regression as a classifier

Design set

		True		
		0	1	Total
Predicted	0	3,086	360	3,446
	1	369	157	526
	Total	3,455	517	3,972

$$\frac{d}{c+d} = 0.30 \ , \ \frac{D}{C+D} = 0.13$$

Test set

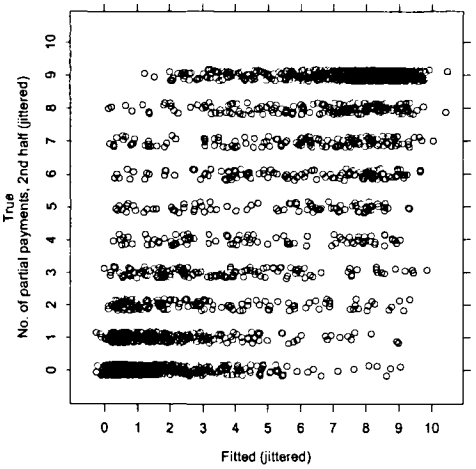
		True		
		0	1	Total
Predicted	0	3,127	333	3,460
	1	349	163	512
	Total	3,476	496	3,972

$$\frac{d}{c+d} = 0.32 \ , \ \frac{D}{C+D} = 0.12$$

In our selected group of predicted intermittent payers, more than 30% really are intermittent payers, compared to around 12% in the overall population - so that mail shots or other incentives can be much more accurately targeted. Given this improvement by using a simple and basic regression approach, we may be able to do even better if we use methods that are more sophisticated. Thus, we can think of the increase in performance from linear regression as the 'baseline' to compare with all other classification methods. Our objective is to find a method that is better than this.

This means that even with this simple technique, on which we hope to improve, we can achieve a targeting accuracy of two and a half times greater than we would have done had we selected a random sample of customers. Including other variables does not improve the fit much – the R^2 increases by less than 0.1. So, we can predict (and accurately) the number of partial repayments that people will make in the future simply by using earlier values of the same variable. However, the inclusion of other variables leads to another problem, illustrated in Figure 5.3.

Figure 5.3 Linear regression, several predictor variables



It is easy to derive our desired summary tables when the only predictor variable is the number of partial repayments, because suitable boundaries for partitioning the data are obvious from Figure 5.2. They are less so if we use other explanatory variables as well, as we show in Figure 5.3 (the other predictor variables in this model were balance, interest paid, delinquency, credit limit and number of cards on the account).

5.1.4 Linear regression, 'real' partial repayments

In Section 5.1.3 we described the results of linear regression when used to model the broader definition of partial repayments (as described in Chapter 4). We expect that the predictions we obtain should be better if we use only the subset of customers who make 'real' partial repayments, rather than including 'forgetful payers' who simply miss payments, as well. This is because the more precise definition ought to result in a group of customers which is more homogeneous. Consider our more rigorous definition of a partial repayer, which we showed in Chapter 4, as follows.

$$x_{t,i} = \begin{cases} 0 \text{ if } \text{balance}_{t-1,i} \leq 0 \\ 1 \text{ if } \left(0 < \left(\text{payment}_{t,i} \div \text{balance}_{t-1,i}\right) < 1\right) \\ 0 \text{ otherwise} \end{cases}$$

Table 5.3 shows how customers group according to this definition, and again, we showed this in Chapter 4. It is similar to Table 4.4, but there are fewer people with '9 partial repayments' in both time periods, and more customers in the 'never make a partial repayment' (group 0) group.

Table 5.3 Number of partial repayments in each 9 month period

First half	Second half 0	1	2	3	4	5	6	7	8	9	Total
0	1,341	144	41	16	8	9	9	14	5	-	1,587
1	153	40	17	18	9	8	8	7	6	6	272
2	52	15	15	14	6	4	8	7	2	14	137
3	25	13	12	11	9	8	11	19	9	8	125
4	19	9	12	12	22	16	12	15	14	8	139
5	19	12	4	16	17	29	26	21	18	23	185
6	9	6	6	12	18	33	37	41	33	28	223
7	11	6	8	13	12	41	46	59	53	54	303
8	5	3	5	9	12	23	38	50	82	85	312
9	5	7	10	6	16	23	34	60	115	413	689
Total	1,639	255	130	127	129	194	229	293	337	639	3,972

Performing linear regressions (with the same explanatory variables as previously) on this more precise definition leads to values of R^2 that are slightly higher than we saw in Section 5.1.3, by around 0.02 – 0.04. This is to be expected, given that the first definition included both 'genuine' partial repayers, and those who appeared in that class simply because their payment arrived a few days late. However, it is not a large improvement, so we conclude that the 'forgetful payers' are having little effect on the overall. We still see a similar number of 'hard core' full payers, and most of the movement has been in the 'occasional' group.

5.1.5 Linear regression, forgetful repayers only

The third type of partial repayers are those customers who do not make a repayment. The models resulting from linear regressions using this definition were much weaker, with R^2 of the order of 0.4. The relationship between the earlier and later times is shown in shown in Table 4.4 that has been re-produced in Table 5.4, and as is obvious, regular forgetful payers are in a small minority. We expect this, because every time a person forgets to pay they are charged interest on the whole balance for the month, even if they are otherwise full payers.

Table 5.4 Number of forgetful payers

First half	Second half 0	1	2	3	4	5	6	7	8	9	Total
0	1,898	351	98	35	11	8	3	-	1	-	2,405
1	319	214	104	46	22	4	2	-	-	-	711
2	98	109	95	64	33	14	3	-	-	1	417
3	31	37	66	40	37	11	8	1	-	-	231
4	16	12	22	22	27	14	5	4	2	-	124
5	5	3	12	12	10	17	4	-	-	-	63
6	3	0	3	1	6	2	2	1	-	-	18
7	-	1	-	1	1	-	-	-	-	-	3
8	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-
Total	2,370	727	400	221	147	70	27	6	3	1	3,972

Most of the reduction in variation explained is because the biggest cell is the one at (0, 0), and, of the other customers, another third only made one late repayment. Any models that seek to fit these data are trying to predict many zeros from many zeros.

5.2 Classical linear discriminant analysis

5.2.1 Introduction

Linear discriminant analysis seeks to allocate each case to a group g, by seeking a linear combination $\mathbf{Ax}$ of the variables which maximises the ratio of the data's between group variance to its within group variance (Mardia et al., 1979). LDA gives the optimal solution if the variables are multivariate normally distributed,

although the method is relatively robust even if this assumption is violated. We have a response variable that has a small number of values, 0 – 9, and it can thus easily be used as a classification variable without any modification. By doing this, of course, we are ignoring the ordinal information implicit in the categories, whereas the regression approach captures this. Using a variety of supervised classification methods on this variable will ensure that we will be able to avoid the problem we faced in Section 5.1, that of having to partition a continuum.

In Table 5.5, we summarise the results from these analyses, where the models shown have independent variables as follows:

LDA, 1 One only - the number of missed payments in the first half of our period.

LDA, 2 The number of missed payments in the first half, plus the following account history variables (24 in all):

MasterCard account held.

Participant in Barclaycard's loyalty scheme.

Number of cards on the account.

Whether the account had been "delinquent" in the first half.

Interest paid, by month.

Ratio of the balance to credit limit, by month.

LDA, 3 Everything as in 'LDA, 2', plus spending in 26 sectors, expressed as a total for the 10 months (they were actually 11 four week periods). There were 40 variables.

LDA, 4 The same variables as in 'LDA, 2', but with spending in 26 sectors, by period (300 variables).

In this section, we discuss the results shown in the first block of five rows in Table 5.5 (the models which used the broader definition of partial repayment), and we discuss the second and third blocks of rows (the 'forgetful payers' and 'real partial repayers') in the next section.

Table 5.5 Summary of classification model results

Model	c	d	d/(c + d)	D/(C + D)
Broader definition				
Linear regression	349	163	0.32	
LDA, 1	349	163	0.32	
LDA, 2	222	169	0.43	0.12
LDA, 3	303	186	0.38	
LDA, 4	323	137	0.30	
Forgetful payers				
Linear regression	616	510	0.45	
LDA, 1	410	312	0.43	
LDA, 2	392	468	0.54	0.28
LDA, 3	428	470	0.52	
LDA, 4	752	593	0.44	
Real partial repayers				
Linear regression	296	115	0.28	
LDA, 1	399	130	0.25	
LDA, 2	474	167	0.26	0.10
LDA, 3	502	165	0.25	
LDA, 4	446	125	0.22	

Table 5.5 also shows the three different definitions of partial repayment that we used in Chapter 4. We expect that, up to some point, the more variables we include, the better will be the performance of our classification rule. We also expect this performance to deteriorate when applied to the test set, and our task is to seek the best compromise of 'good' performance in the first instance, but with little deterioration when applied to the test set.

Looking at Table 5.5, we see varying $d/(c + d)$ ratios with different values of $(c + d)$. With $(c + d)$ around 500 we get a $d/(c + d)$ ratio of around 30%, but with $(c + d)$ around 400 we get 40%. There are two points to be considered: (i) whether $(c + d)$ of 400 is too small to be useful (as it is only around 10% of the overall test set) and (ii) the fact that $d/(c + d)$ is inversely related to $(c + d)$. Taking (i) first: 10% of

Barclaycard's customer file is more than half a million individuals, so anything that can improve the targeting among such large group is worthwhile. Secondly, the relationship described in point (ii) becomes important if a model were to have a large $(c + d)$, but $c >> d$. For a model which performs as well as the ones we have just outlined, a large $(c + d)$ would not be necessary.

The degradation in performance that we see in the 'LDA, 4' model might be expected, following application of the classification rule to the test set. With the use of spending by trade sector, by month, we added another 250 plus variables to the analysis, and this is reflected in the better performance in the design set (as the rule is fitting more randomness), leading to poorer performance in the test set.

Generally, we are looking for a combination of predictive power, but ideally combined with parsimony, providing those two aims are not mutually exclusive. A model with fewer variables to be collected, cleaned and checked is more likely to be implemented quickly and with fewer faults. This means that, based on classical LDA, we would select the second model of those just shown. This is the model labelled 'LDA, 2' in Table 5.5, and it contained account history variables, but none that were of spending, either in total, by trade sector, or by time. It performed three and a half times better than we would achieve by selecting a random sample of customers; and there was little deterioration when it was applied to the test set. In each case, d is large enough to be worth investigating further.

5.2.2 More precise definitions of partial or forgetful payers

All of the LDAs that we discussed above were based on our broader definition of a partial repayer. We now describe the results obtained from the two main subgroups

of that broad definition – the second and third blocks of rows in Table 5.5 – as we did for linear regression in Sections 5.1.4 and 5.1.5. A priori, we would expect classification performance to be better with these groups, because they are defined more precisely than the broader definition.

The models we built for these two groups used independent variables as before, with the dependent variable taken from the second half of our period. It is, for the two more precisely defined classes, the number of forgetful payers, and the number of 'real partial repayers'. We will not describe the results from linear regressions in which we used the more precise definitions of partial repayment because, as it did then, the simpler approach did not give such good results as discriminant analysis.

We see from Table 5.5 that all models performed better than they would have done had we selected a random sample of customers, and all would appear to be worth pursuing. Note that the results from linear regression and the first LDA model (for our broader definition) in Table 5.5 appear to be the same, but this is only because we have summarised the predictions, which were into ten groups, into our 2×2 confusion matrix. If we had showed a 10×10 matrix we would have seen that the results were slightly different.

5.2.3 Conclusions

Simple linear regression gave better predictions than we would have achieved if we had selected a random sample of customers, but the method gave rise to another consideration – that of partitioning the space for classification. Also, the classifications we obtained from regression did not perform as well as LDA, when assessed against the measure we had defined. As this assessment method was driven

by the business' needs, this was an important criterion for selecting LDA in preference to regression.

We saw that some simpler LDA models, with few independent variables, gave better results than those with a greater number of variables. Moreover, the additional variables that led to a poorer fit were always spending (as distinct from 'account history') variables. We mention this fact to highlight that repayment is made against the total balance outstanding, while spending decisions can be made independently. Barclaycard knows, from qualitative research, that consumers have a tendency to separate these two aspects of behaviour. They can thus use their cards for the pleasurable (spending), while being able to forget that the unpleasant (repayment) must be made at some time. All definitions of partial or late repayment gave similar improvements in predictive performance.

These results are encouraging, although there are some areas that need further investigation. From a practical point of view, as far as the company is concerned, improved targeting is already possible with them. Not only that, but there could be immediate financial gains, even if only a small number of customers in our target group responded. In some senses, because the models only predicted a small number, around 10% of the sample, into our 'suitable for targeting' group, we might draw the conclusion that these results are not promising. However, this 10%, when applied to the whole file, is considerably more than would commonly be used for test marketing activities. As a business application, we believe that we have devised a rule that could be useful for incentivising groups of customers. In fact, since this work commenced, Barclaycard has started to carry out much of this sort of activity, and often for the first time. In Section 5.3 we explore these ideas further, and

examine alternative kinds of predictive model. LDA also has one useful property as far as our data are concerned – it is reasonably robust to non-normality, which is apparent from the consistency of performance in the rules that we have described.

5.3 Separation of the classes

We have now performed several analyses that have been, to some extent, successful. We mean by this that we could use them as practical discrimination tools. Certain features of the results, though, lead us to believe that the classes are not well separated, and these are summarised as follows. To examine this we use classical LDA, which is one of many techniques we could use (some are discussed in Section 5.5) to produce some low dimensional plots that will enable us to visualise the overlaps between groups.

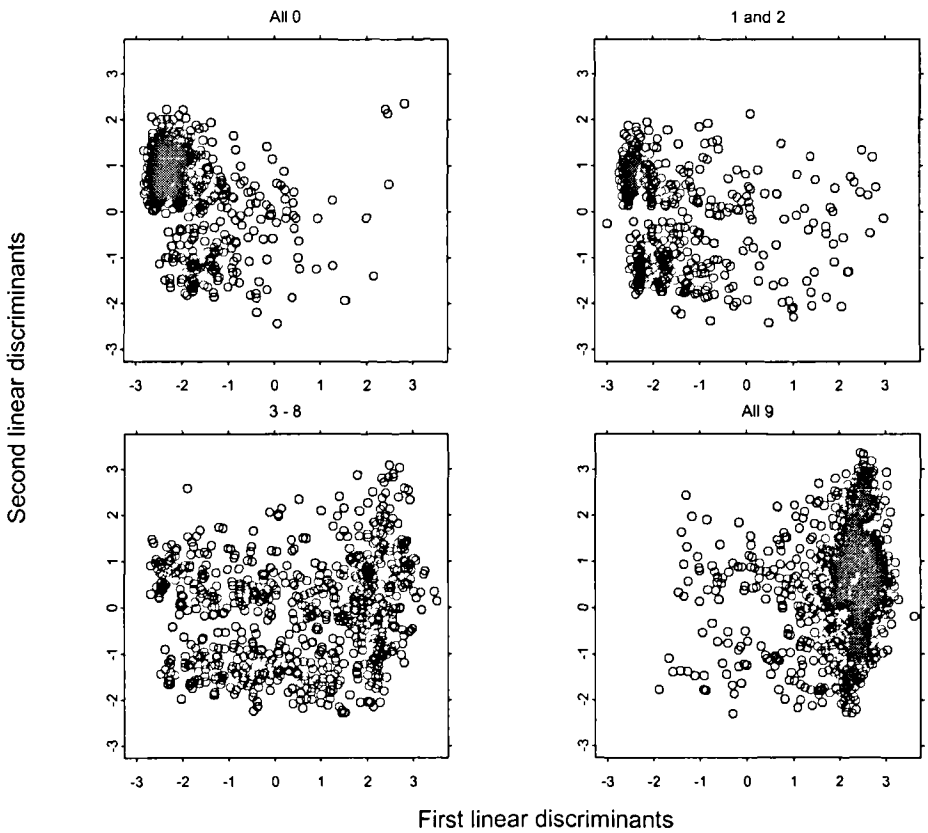
5.3.1 Discriminant analysis plots to investigate separability

The distributions of the number of partial repayments (as shown, for example, in Table 4.1) illustrate that while the extremes seem to be well separated, the other classes might be overlapping, to quite a large extent, in the space of predictor variables. In Figure 5.4 we show the extent of this overlap. It shows the first two linear discriminants, but to see where the different points lie, we have separated them onto four separate panels as follows: no partial repayments ('All 0'), 1 or 2 partial repayments, who are the customers we think we will be able to incentivise ('1 and 2'), three to eight partial repayments ('3 – 8'), and nine ('All 9'), and partial repayments. As is clearly seen, the extremes of 0 and 9 are on the left and right hand sides of the plot respectively, where they are densely packed. The 1s and 2s tend to be towards the left, while the remainder of customers are spread more or less evenly,

and there is a gradual shift in the density of people, by their number of partial repayments, with the lower numbers concentrated towards the left, and the higher ones to the right.

The problem now becomes readily apparent, the group of interest – '1 and 2' – appears to almost to overlap with the full payers. Plotting any other two of the linear discriminants shows similar characteristics.

Figure 5.4 Broader definition, first two linear discriminants



5.3.2 Separability - summary

With the variables that we are using, the people with one or two partial repayments have a considerable degree of overlap with one of the extreme groups, zero partial repayments.

We defined, in Chapter 4, two other ways to define partial repayers. We will not show the plots of their discriminant functions, because they had similar characteristics, at least for the problem we have – the groups we defined to be of interest were largely co-located with the full payers. Thus, there is likely to be little further improvement, given this overlap.

5.4 Nearest neighbour methods

5.4.1 Introduction

Our classes overlap to a certain extent, as we showed in Figure 5.4, which means that any classification method could have problems when trying to separate the groups of interest. A method that does not impose a linear decision threshold could be more successful than LDA, and one which uses a local model, as *k*-nn does, may allow us to discriminate our group from the extremes of the data.

5.4.2 Choice of *k*

In our work on LDA, we described two broad approaches to classification. The first was to predict membership of all ten classes of partial repayments, 0, ..., 9, then to summarise the results in a confusion matrix, which we showed in Table 5.1. The second approach was to group customers into the 'of interest / not of interest' groups before trying to predict class membership, and then to predict membership of these two classes. Application of *k*-nn to our problem will vary according to which of

these prior groupings, two or ten, we seek to predict class membership. We will not give a detailed explanation of the models we produced, but will instead summarise the problems we encountered with the method.

Choice of k is crucial in any k-nn analysis, and it involves a trade off between bias if k is too large, and variance if k is too small. Henley and Hand (1997) describe formal methods for the selection of k.

In the ten class case, for commercial reasons, we selected k based on comparison of the ratios $d/(c+d)$ and $D/(C+D)$ at several values of k. We chose the predicted class to be the one that had the greatest number of points amongst the k nearest neighbours.

In the two class case, we needed to consider the threshold with which to compare the k-nn estimate, an issue discussed in Adams and Hand (1999), who give the threshold that will result in minimum future loss. It is the value t such that the object is classified into class 0 whenever $\hat{p}(0|x) > t$, where $\hat{p} = \hat{p}(0|x)$ is the estimated probability that an object with measurement vector x will belong to class 0. Adams and Hand (1999) then state that 'minimum loss is achieved by choosing the classification threshold such that points are classified into class 0 if $\hat{p} > t = c_1/(c_0 + c_1)$'. This is based on the relative costs of the two kinds of misclassification, as shown in Table 5.6.

Table 5.6 Costs of misclassification

		True		
		0	1	Total
Predicted	0	0	c_1	
	1	c_0	0	
	Total			

In our case, however, c_1 is not a cost at all, but it might be thought of as the 'lost opportunity cost'. If the business does not incentivise any customers, it costs nothing to do so, at least in the short term. If it *does* incentivise different parts of the file, some of the customers to whom we offer an incentive will generate extra revenue for the business, and our threshold must incorporate this fact too. We will cover this in more detail in Chapter 6, but essentially the table we need to use is shown in Table 5.7.

Table 5.7 Costs of misclassification

		True		
		0	1	Total
Predicted	0	0	c_1	
	1	c_0	c_2	
	Total			

Where, in our case, the 'costs' can be considered as follows.

c_0 - Cost of misclassifying a class 0 customer, or, in our case, or the cost of mailing those customers we predict to be in the group of interest, but who are not.

c_1 - Cost of misclassifying a class 1 customer, the lost revenue because we did not mail those customers who are likely to respond to our incentive.

c_2 - Cost generated by those customers who respond to our incentive, incorporating the revenue they generate.

Thus, the total cost is $c_0 + c_1 + c_2$, and if the result of this sum is negative, our incentive has generated revenue for the business, rather than cost. Adams and Hand's (1999) assumption of no cost being attributable to correctly classified objects does not apply here, so we cannot select our threshold using the simple relationship they showed.

We will investigate revenues and costs in more detail in Chapter 6, but the revenue from any activity will be as follows.

$$\text{Net revenue} = -(c_0 + c_1 + c_2) \tag{1}$$

Here, the elements of equation (1) are as follows, where i is the cost per item mailed, r is the revenue generated by an individual who responds, s is the response rate (i.e. the proportion of people in the target group who take up the offer) and b, c and d are as shown in Table 5.1.

$$c_0 = ci, \quad c_1 = bi - brs, \quad c_2 = di - drs \tag{2}$$

However, c_1 is a notional cost incurred because we do not mail customers in this group – they are not predicted to be in our group of interest. Our objective is to maximise net revenue (or to minimise net cost), and (1) and (2) can be combined to give the following.

$$\text{Net revenue} = -(ci + di - drs)$$

Or,

$$\text{Net revenue} = d(rs - i) - ci$$

Of these, only r is known or can be estimated from the data, while b, c and d depend on the predictive rules we calculate. The cost per item, i, depends on the number of customers we choose to mail.

5.4.3 Relationship between k and the threshold

In our case, only 13% of customers are in the group of interest, and we might expect our classification threshold to be different from 0.5 because of this (if $c_0 \cong c_1$ and $c_2 = 0$ the threshold t will be approximately 0.5). The revenue per customer who responds is considerably greater than the cost of mailing to that individual (we cover this in more detail in Chapter 6). If t is considerably less than 0.5, we will see an interesting phenomenon as a consequence. Assume, for the purposes of illustration, $t = 0.125$. If $k = 1, 2, \ldots, 8$ is used, then any customer who has just 1 out of the k nearest neighbours belonging to class 1 will be classified into class 1. However, if $k = 9$, then any customer who has just 1 out of the k nearest neighbours belonging to class 1 will be classified into class 0. A similar jump in predictions will occur after each $k = 8, 16, 24, \ldots,$. This switching also leads to an interesting property of the k-nn method – that the choice of k must depend on the threshold, which itself is dependent on costs.

5.4.4 Choice of a suitable distance metric

It is well known that the effectiveness of nearest neighbour methods depends critically on the choice of distance measure. A common metric is the Euclidean distance, but one of the potential pitfalls of using this is that all of the variables are, by default, treated with equal importance, and it may also subject to an arbitrary choice of measurement unit for non-commensurate variables. A common standardisation is to subtract the mean from each variable, then divide each by its

standard deviation. This, however, imposes a notion of relative importance that may have little to do with their relative importance for discriminatory purposes.

We used the methods described in Henley and Hand (1996, 1997) that would allow us to learn from information in the data, and to replace the standard Euclidean metric

$$d = \{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})\}^{\frac{1}{2}}$$

with an adjusted metric

$$d = \{(\mathbf{x}-\mathbf{y})^T(\alpha\mathbf{I}+(1-\alpha)\mathbf{w}\mathbf{w}^T)(\mathbf{x}-\mathbf{y})\}^{\frac{1}{2}}$$

We changed their matrix $(\mathbf{I}+D\mathbf{w}\mathbf{w}^T)$ to $(\alpha\mathbf{I}+(1-\alpha)\mathbf{w}\mathbf{w}^T)$, which effectively allowed us to choose values of D that could vary between zero and infinity. The variable $\mathbf{w}$ is the vector of coefficients from a standard linear regression, where the variables are those used in the k-nn analysis. This approach seeks to improve the distance metric by incorporating knowledge from the data, and our metric includes a parameter α, which needs to be chosen (as $\alpha \to 1$ the metric tends towards the Euclidean distance).

The results from models where we experimented with values of α between 0.001 and 0.99 were similar to those in which we used the simple Euclidean metric. In most cases, using Euclidean distance, or some standardised or adjusted variant, k-nn resulted in poorer performance than LDA.

5.4.5 Conclusions: *k*-nearest neighbour methods

In our investigation of k-nearest neighbour methods, two distinct and important issues arose. For classification into ten classes, we found that the ratio $d/(c+d)$

improved, but $(c + d)$ decreased with increasing k, which resulted in smaller populations for targeting. Further, as k increased, more and more customers were classified into class 0 or 9, which are the two extremes – the 'never' and 'always' make a partial repayment groups. We suspect this arises because our groups of interest are a (relatively) small proportion of the sample. We have concentrated on customers who made one or two partial repayments, because we believe they are most likely to respond to incentives. Intermittent partial repayers are valuable to the business, whichever definition of partial repayment we choose to use, but they are spread between the extremes and overlap, to a certain extent, with customers who made no partial repayments. This is why, with increasing k, we quickly reached a situation (typically as $k \rightarrow 30$) where the greatest number of nearest neighbours would be in one of these extremes, usually group 0.

For the case where we grouped the ten classes into two groups *before* classification, we had to make a choice about the most appropriate threshold that would minimise future expected loss. With little illumination to be gained from the costs of misclassification, we found that the optimum threshold was close to the sample prior proportion. Even at this optimum, the best value of the $d/(c + d)$ ratio was almost the same as the one that resulted from the use of LDA.

We experimented with different distance metrics, by adapting them to learn from the data, using a formula that would allow us to vary the importance given to the regression coefficients by varying the parameter α. The results at any of the values we selected for α resulted in classification performance that was little different from the ones in which we had used the Euclidean metric. We also investigated the effect

of standardising the variables, but that produced similar results to those of the Euclidean metric.

In most cases, *k*-nn models gave worse results than we had, a priori, expected. Most customers (in our groups of interest) are closer to more cases in the extremes than customers in the same class. This effect is exacerbated because the two groups of interest make up around one eighth of the total, but the people with 0 and 9 partial repayments are each around a third. This is probably why, as *k* increased, most accounts were allocated to one of the extreme groups. Another way of saying this is that cases in group 0 swamped any local neighbourhood as k increased. Changing the classification threshold *t* in the two class case led to little improvement, but introduced a new problem – that of how to select the optimum value for *t*.

5.5 Other types of discriminant analysis

There are many other variants of discriminant analysis we could use. We have investigated regression methods, linear discriminant analysis, and nearest neighbour methods for classifying the partial repayers. However, there are still more methods, including quadratic and logistic discrimination, recursive partitioning methods, neural networks, projection pursuit regression, kernel methods, support vector machines, and an almost infinite number of variants of these. There are also many general books on the topic, including in Hand (1981), Devijver and Kittler (1982), McLachlan (1992), Ripley (1996), Hand (1997), and Webb (1999). We based our choice on grounds of simplicity and ease of understanding, so that the methods we explore can be applied by non-experts should they prove to be effective.

5.6 Conclusions

Simple regression gave better predictions than the selection of a random sample of customers would have done, but if we used more than one predictor variable the technique led to another problem (although this can be overcome) – that of partitioning the space for classification. We saw an improvement in performance with classical LDA, despite the classes being poorly separated, and we developed predictive rules that resulted in large enough groups to be used for targeting purposes. We were able to improve predictive power by a factor of three compared to the file as a whole. Finally, we saw that classical linear discriminant analysis worked considerably better than *k*-nn methods. Several features were apparent from the models we produced, and we summarise them below.

1. Most of our classifications showed some sort of separation.
2. This separation was usually at the extremes of the distributions, with the 0s and 9s well separated.
3. The 0s were more densely packed than the 9s.
4. In between the two extremes, there was a much more diffuse distribution, which contained customers from all groups.

The results we saw in Section 5.3 led us to believe that our groups overlapped to some extent. None of the better performing models contained any information about customers' spending, and we speculated that this was because of the separation, in consumers' minds, of their repayment and purchasing behaviour.

We ought to expect poor discrimination, since we are trying to separate out a segment from a continuum (the number of partial repayments) and we would not really expect to be able to predict this continuum with great accuracy. We are not

really interested in 'predictive accuracy' as an end point, but rather as a means to an end, in this case as a means to a commercial end - and it does seem that the accuracy is sufficient for it to be valuable for commercial purposes.

From a practical point of view, these results are useful immediately, and there will be financial benefits, even if only a small number of customers in our target group respond. We achieved a threefold improvement in prediction, compared to the sample as a whole, with little degradation when applied to the test set. So, we believe we have devised a rule that can be used for incentivising groups of customers. Commercial sensitivities mean that we cannot reveal exactly how this would be implemented, or how many customers would be more likely to respond, but it is around ten times more than the minimum number of customers Barclaycard would usually seek to incentivise.

In this chapter and the previous one, we have made two important discoveries: (i) the fact that, using interpolation between calculated values, we can calculate the long term probabilities of partial repayment, missed payment, and full payment, and (ii) the fact that we can build predictive rules of sufficient accuracy to be commercially interesting. An alternative approach to the problem might be a Markov transition model, explored in such work as Cyert et al. (1962), Kallberg and Saunders (1983), Frydman et al. (1985) and Till (2001).

In Chapter 6, we show how the methods we have described can be used to generate value for the business.

Chapter 6

6 Profitability gains from discriminant analyses

6.1 Cost-benefit analysis

We showed a contingency table in Chapter 5 that we used for assessing the classification the rules we developed, and we repeat it in Table 6.1.

Table 6.1 The general assessment table

		True		
		0	1	Total
Predicted	0	*a*	*b*	
	1	*c*	*d*	
	Total	*C*	*D*	

Customers in class 1 (in Table 6.1) are of most interest, because we believe we can offer them incentives that will change their behaviour, and they are the customers with one or two partial repayments. All other customers in this table are in class 0. We now describe an equation that will allow us to calculate whether a mailing to our predicted group is of financial benefit to the business. As before, we are interested in $c + d$ and $d/(c + d)$, and we will use the same notation in this chapter. We showed the costs of misclassification in Chapter 5, and we repeat them in Table 6.2. Our approach differs from that of Adams and Hand (1999), because our matrix has an extra entry, c_2, where theirs had a zero.

Table 6.2 Costs of misclassification

		True		
		0	1	Total
Predicted	0	0	c_1	
	1	c_0	c_2	
	Total			

For a proper assessment of return from an incentive, we need the cost matrix to have four non-zero entries, as we show in Table 6.3. In this table, c_3 is the cost of mailing those customers whom we correctly predict to be in the 'not of interest for incentivisation' group. It might seem odd that we need to do this, but, as we shall demonstrate, it is an important part of assessing the effectiveness of our predictive rules. We further show that the second part of the assessment reduces to a comparison of the $d/(c + d)$ and $D/(C + D)$ ratios.

Table 6.3 Costs of misclassification

		True		
		0	1	Total
Predicted	0	c_3	c_1	
	1	c_0	c_2	
	Total			

6.1.1 A general equation for assessing profitability

For most of the results we gave in Chapter 5, we showed a, b, c and d as numbers of customers, but we can re-express Table 6.1 to show proportions instead: see Table 6.4. This allows us to develop a general equation for assessing the profitability of any marketing activity.

Table 6.4 Proportions of real partial repayers

Test set		True		
		0	1	Total
Predicted	0	p_a	p_b	
	1	p_c	p_d	
	Total			

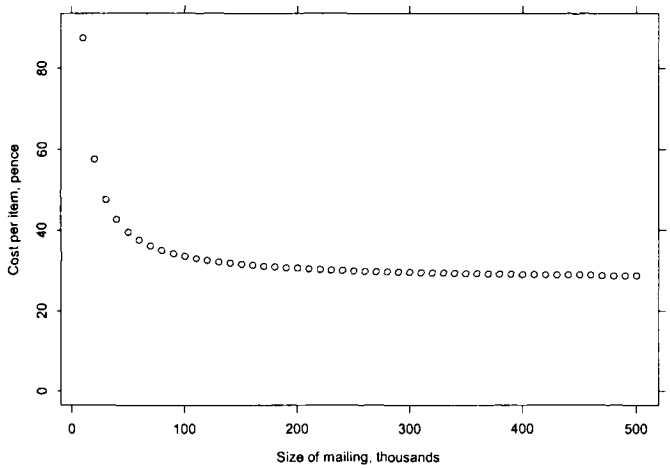
Here $p_a + p_b + p_c + p_d = 1$. In the more usual notation for this type of table, we would have $p_a + p_c = \pi_0$, $p_b + p_d = \pi_1$, $p_a + p_b = p_0$ and $p_c + p_d = p_1$. Where π_0 is the prior probability of class 0, $\pi_1 = 1 - \pi_0$, p_0 is the proportion predicted to have come from class 0 and $p_1 = 1 - p_0$.

6.1.2 Analysis of revenue and costs

Usually, we do not know the value of c_1, so we must make a two stage assessment of the success of an incentive. The first is to examine the costs and revenue from that incentive alone, and the second is to compare it with the performance we would achieve if we mailed a random sample of customers from the whole file.

The cost per item for marketing activity depends on the number of people we choose to mail, and we show the relationship between size of mailing and cost per item in Figure 6.1. So, for any mailing to m customers we know the cost of each mail pack.

Figure 6.1 Cost per item, and size of mailing



In Chapter 5, we showed that net revenue was

$$-(c_0 + c_1 + c_2)$$

where $c_0 = ci$, $c_1 = bi - brs$, $c_2 = di - drs$, i is the cost per item mailed, r is the revenue generated by those who respond, s is the response rate, and b, c and d are as shown in Table 6.1. Here, for assessing the performance of the incentive, and, as described in Chapter 5, ignoring the notional cost c_1, we have the following.

$$\text{Net revenue} = -(c_0 + c_2)$$

We rarely incentivise the whole file, so let m be the number of customers we choose to mail for any particular activity. Here, because we have expressed the tables as proportions, we make the following substitutions.

$$ci \text{ becomes } \frac{p_c}{p_c + p_d} im, \quad \text{and} \quad di - drs \text{ becomes } \frac{p_d}{p_c + p_d} im - \frac{p_d}{p_c + p_d} mrs$$

$$\text{Net revenue} = -\left(\frac{p_c}{p_c + p_d} im + \frac{p_d}{p_c + p_d} im - \frac{p_d}{p_c + p_d} mrs\right)$$

Simplifying this gives

$$\text{Net revenue} = \frac{p_d}{p_c + p_d} mrs - im \qquad (1)$$

6.1.3 Real partial repayers results

Consider the results from the model 'real partial repayers, LDA, model 2', which we showed in Table 5.5 – expressing the number of customers as proportions results in Table 6.5.

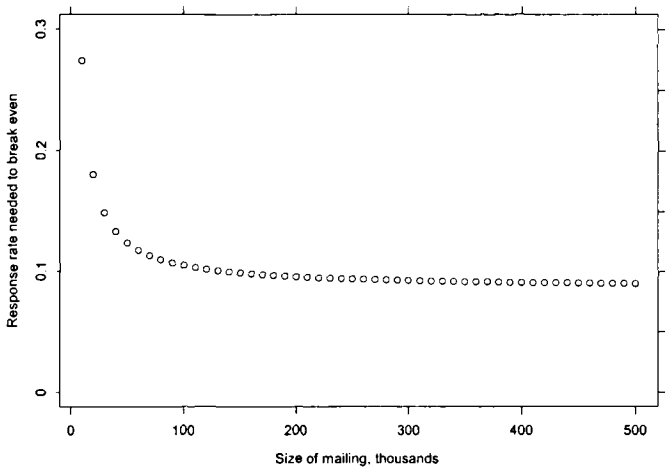
Table 6.5 'Real' partial repayers

Test set		True		
		0	1	Total
Predicted	0	0.776	0.062	0.839
	1	0.119	0.042	0.161
	Total	0.896	0.104	1

$$\frac{d}{c+d} = 0.261 \ , \ \frac{D}{C+D} = 0.104$$

We can now calculate, for any given mailing to m customers in our predicted group, the response rate necessary to break even, and we show this in Figure 6.2, using equation (1). The points show where net revenue is zero, we have estimated r from the data, and i is known for each value of m.

Figure 6.2 Response rate needed to break even – real partial repayers



We can now assess the response rate necessary to achieve break even from any particular incentive. We will not show the profitability calculations for any of the other models described in Chapter 5, but the method works in a similar way for each of them. We have thus derived a general equation that can assess the financial returns from any incentive, and it will allow us to determine the response rate at which an initiative breaks even.

6.1.4 A random sample of customers from the whole file

The second stage of the assessment is to compare a mailing of size m to customers in our predicted group with a mailing of size m to a sample of customers from the whole file. We do not know c_1, but we know the sum of the proportions $p_a + p_b = p_0$ so we can calculate $c_1 + c_3$. Thus, we have the following relationship for net revenue if we mail a random sample of customers from the whole file.

$$\text{Net revenue} = (p_b + p_d)mrs - im \qquad (2)$$

Comparing equation (2) with equation (1), we see that a mailing to m customers in our predicted group will lead to greater revenue than a mailing to m customers selected at random when the following is true.

$$\frac{p_d}{p_c + p_d} > p_b + p_d$$

This is (effectively) the same as the comparison of the ratios $d/(c + d)$ and $D/(C + D)$, and it illustrates another reason why this is the most important way – for the business – of assessing the effectiveness of our rules. We can now assess, and easily, the merits of targeting particular groups of customers.

We have been rather harsh in assessing our results, because we have assumed that customers who are not in our target group will not respond to incentives. In reality, this is unlikely to be the case, but the response rate from these customers is likely to be much lower than among the group of interest.

6.1.5 *k*-nearest neighbour methods

We have not mentioned *k*-nn methods so far in this chapter. They often performed better than the simpler classical LDA methods, in that the ratio $d/(c + d)$ was often high, especially at relatively large k (say $k > 30$, for our data), when compared to classical LDA. However, if $d/(c + d)$ *was* high, the sum $c + d$ was usually considerably lower than in comparable LDA models, and the *k*-nn analysis might have found a target group of only 100,000 customers in the whole file. Also, we could not improve the method by selecting different classification thresholds. The profitability calculations reinforce our view that, from a practical (i.e. business) point of view, the simpler method is considerably superior to the more flexible approach.

Finally, the discussion in Chapter 5 about time and complexity of the method would lead us to question their use in a fast moving business environment.

6.2 Conclusions

We have shown a simple equation for determining the profitability of a campaign, where most of the variables are either known, or can be set by the business. We know, from Chapters 4 and 5, that the difficulty is in determining c and d in Table 6.1. In those chapters, we saw that the simpler linear methods worked as well as k-nearest neighbour (from the business' point of view) because they led to larger groups for targeting purposes. The equation we developed allows easy comparison between different types of incentives to different types of customer groups.

Also, we can assess the effectiveness of a single campaign, but we have to do so in two stages. The first stage is to ignore the notional costs, and concentrate only on the revenues and costs that can be attributed directly to a particular mailing campaign. The second stage is an assessment very similar to the one we described in Chapter 5 – comparison of the ratios $d/(c + d)$ and $D/(C + D)$.

Chapter 7

7 Transaction data – descriptive models

7.1 Introduction

A theme that runs through our work is that of recognising and then modelling behaviour. In Chapters 4-6, we described models of repayment behaviour, and these were based on data we called 'account history'. A distinguishing feature of this type of data is the fixed number of records, usually based on a monthly summary of behaviour, some of which is recorded on customers' statements. In the case of credit cards, which we describe, data are captured at the time the monthly statement is produced. In the case of loans, it will often be when the monthly payment is due. Much work has been done over the years on the use of these types of data for the prediction of bad debt. See, for example, Hand and Henley (1997) or Hand and Jacka (1998) for details of different methods.

In Chapter 3, we described how, in the 'account history' part of our data, the transactions a customer makes each month are recorded as a single figure, which is the total amount spent in the previous month. In 2000, there were 1.5 billion transactions on UK issued credit cards (BBA, 2001), in almost a thousand individual merchant categories, and across 15 million merchants in the UK and around the world. In this chapter, we describe some of the ways that credit cards are used for spending. We show that the data, although by their nature much more variable than account history, have a great deal of structure, if they are examined in an appropriate way.

Initially, we will describe our modelling in only one sector – petrol stations – and, in Chapter 8, we will describe the fitting of univariate distributions to each sector. Then, we will outline our attempts to find ways to describe the links between the use of different trade sectors. There are various ways to do this – number of transactions, value of these transactions, number of sectors used and so on and we will describe several alternative approaches for these.

Most of the data we use are based on the sample of 3,972 accounts we described in Chapter 3, and we describe customers' spending behaviour in 1996. At this stage, we have restricted ourselves to one year, and Chapter 9 will describe our attempts to model these data and predict behaviour in the following year.

7.2 Summary information

First, we describe how we have summarised individual 'Merchant Category Codes' (MCCs) into 26 more homogeneous trade sectors, and in Table 7.1 we give some summary information for these. We have chosen these groupings with the aim of producing homogeneity, although the 'other' sectors will always be more heterogeneous than most others (such as supermarkets, for example) will. Note that 'others' contain MCCs which do not easily fit into one of the named sectors, but which can contain many individual codes. To illustrate this, our 'other' sector contains almost 150 individual MCCs. Two typical examples of small MCCs are vets and costume parlours, which each account for less than 0.1% of total turnover. We will see, however, one of the drawbacks of such aggregation several times in the coming chapters. For example, 'other shop' and 'other' sectors are the most frequently used, but this appears to be because of the variety of different types of outlets they contain – such as *duty free shops*, *discount stores* and *hearing aid stores*.

If we were searching for patterns, as distinct from models, a difference we described in Chapter 1, then we might need to use data at a less coarse level of aggregation.

Wherever possible, the sectors are as homogeneous as we can make them. For example, the sector 'supermarkets' contains those MCCs which are supermarkets, and nothing else; 'clothing' contains *boys' clothing*, *men's and boys' clothing*, *women's clothing*, *clothing - speciality shops*, *women's accessory and speciality shops*, *children's and infants wear stores*, *family clothing*, *sports and riding apparel*, and *men's and women's clothing stores*. In many cases, use of these will be similar, and it makes the problem of analysing spending behaviour more manageable. Using the individual MCCs could also lead to problems, because of the different 'levels' of data they conceal. For example, we might expect the profile of Waitrose shoppers to be different from those of Kwik Save, yet our data do not allow us to classify at that level. For hotels, airlines and car hire companies – which have individual MCCs - we could do so, and if we were to use such data, we might need to consider weighting the variables.

MCC data would be crucial if we were seeking to identify the differences between travellers on airline A and airline B, but that is not our objective here. With data available at the time of writing, rather than when we took our extract, it is now possible to look at spending in different supermarkets, for example. A specific database has now been created which allows this, among other things.

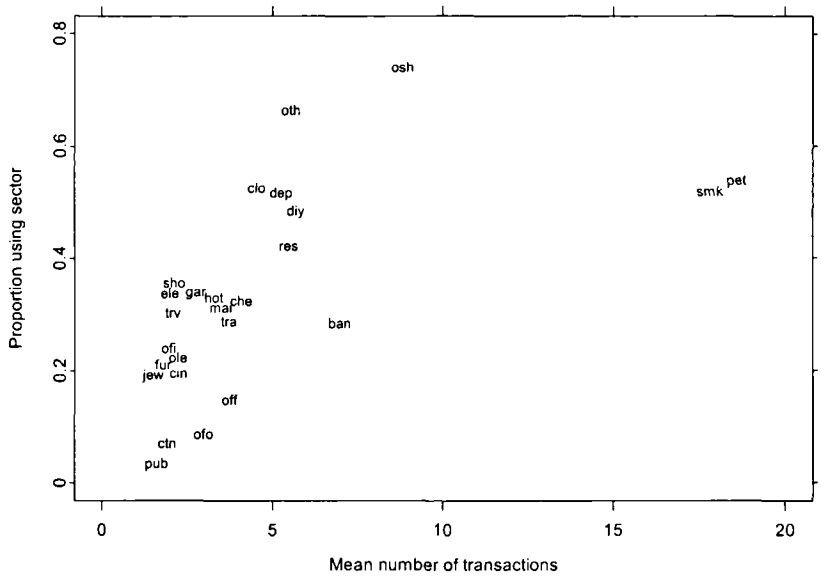
Table 7.1 Trade sector summary

Sector	Proportion of the total amount spent (1)	Proportion of the total number of transactions (2)	Proportion of accounts using the sector (3)	Transactions per account, per sector (4)	Mean transaction size (5)	Median transaction size (6)
Cash advances	7.4%	3.8%	28.8%	7.0	£92.31	£50.00
Chemist	1.0%	2.5%	32.7%	4.1	£18.14	£13.25
Cinema	0.7%	0.9%	21.2%	2.1	£40.16	£22.50
Clothing	4.4%	4.6%	52.9%	4.5	£45.84	£28.00
CTN	0.2%	0.3%	7.5%	1.9	£26.69	£16.55
Department store	4.4%	5.2%	52.0%	5.3	£40.34	£24.00
DIY	3.9%	5.3%	48.8%	5.7	£35.19	£19.99
Electrical goods	3.0%	1.3%	34.1%	2.0	£108.05	£39.99
Furniture	2.6%	0.8%	21.4%	2.0	£150.78	£57.50
Garage	3.3%	1.8%	34.4%	2.7	£85.97	£41.63
Hotel	4.6%	2.1%	31.9%	3.5	£103.82	£60.95
Jeweller	0.9%	0.6%	19.6%	1.5	£72.40	£33.50
Mail order	2.9%	2.1%	31.4%	3.5	£66.50	£24.00
Off licence	0.5%	1.1%	15.1%	3.7	£23.19	£14.46
Other	8.5%	7.0%	66.8%	5.5	£57.26	£24.23
Other finance	2.6%	0.9%	24.3%	2.0	£137.89	£97.13
Other food	0.3%	0.5%	9.0%	3.0	£24.25	£17.22
Other leisure	0.8%	0.9%	21.3%	2.1	£43.76	£23.40
Other shop	10.1%	12.5%	74.5%	8.8	£38.23	£20.64
Petrol stations	8.2%	19.2%	54.4%	18.6	£20.26	£18.50
Public houses	0.1%	0.1%	3.9%	1.6	£44.67	£28.00
Restaurant	3.5%	4.4%	42.4%	5.5	£37.75	£27.55
Shoe shop	1.1%	1.4%	34.4%	2.1	£36.29	£28.88
Supermarket	13.4%	17.7%	52.3%	17.8	£35.84	£26.33
Transport	3.0%	2.1%	29.0%	3.7	£69.63	£29.70
Travel agent	8.6%	1.2%	30.6%	2.1	£333.04	£167.80

Note that the figures in columns 4, 5 and 6 are based only on those people who actually used the sector, not all customers. Even at this simple level, it can be seen that there are differences between different sectors. It should also be apparent from the difference between the mean and the median transaction size that each sector's distribution is skewed, and we investigate this more fully in Chapter 8.

Figure 7.1 shows the data from columns (3) and (4) in Table 7.1, with the sector names abbreviated with the labels shown in Table 7.2. Note that, because of over-plotting, we have moved five points, but only slightly.

Figure 7.1 Mean number of transactions per account per sector



There are some interesting features about this plot. Firstly, most of the sectors lie close to the line $y = 0.083x + 0.038$, but with three exceptions – cash advances, supermarkets and petrol stations. We will mention these three throughout this, and the next two, chapters, because they are obviously different from most others. Supermarkets and petrol stations because of the high number of transactions in these sectors, and cash advances because every analysis we undertake has cash advances as a distinct point.

Further, some sectors appear in groups (e.g. 'jew', jeweller, 'fur', furniture, 'cin', cinema, 'ofi', other finance, 'ole', other leisure), and we will return to this in Chapter 8, where we develop a taxonomy of sector use. Other analyses shown there demonstrate why we have selected these particular groupings.

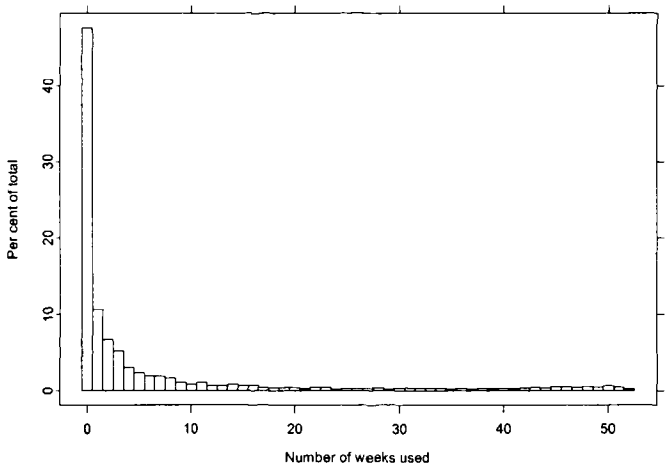
Table 7.2 Abbreviations for sector names

ban	Bank (cash)	gar	Garage	smk	Supermarket
che	Chemist	hot	Hotel	tra	Transport
cin	Cinema	jew	Jeweller	trv	Travel agent
clo	Clothing shop	mai	Mail order	ofo	Other food
ctn	CTN	off	Off licence	osh	Other shop
dep	Dept. store	pet	Petrol station	ole	Other leisure
diy	DIY	pub	Public houses	ofi	Other finance
ele	Electrical	res	Restaurant	oth	Other
fur	Furniture	sho	Shoe shop		

7.3 Features of these data

It might seem that with the richness of these data, we know a lot about how Barclaycard's customers use their cards for spending. People in our design set made more than 200,000 transactions in 1996, and customers in the whole sample made more than 1,000,000 transactions in the two years we had available. This is true, but credit cards are just one of the many ways by which customers can purchase goods and services. This means that our data probably contain only a minority of customers' spending – APACS (2001c) estimates that credit card holders use their credit cards for less than a sixth of their total spending. We illustrate one manifestation of this in Figure 7.2, where it is apparent that almost half of Barclaycard's customers do not use their cards in a supermarket, because of the height of the peak at zero, but presumably most people eat, and regularly. We conclude that many people use debit cards, cheques or cash rather than their Barclaycard in supermarkets, and there will be customers who never shop in supermarkets. Barclaycard is only one of the many credit card brands in the U. K. (albeit the biggest), but our data set contains information only on Barclaycard activity, not any of the other credit cards in circulation.

Figure 7.2 Number of weeks per year that customers use their cards in supermarkets

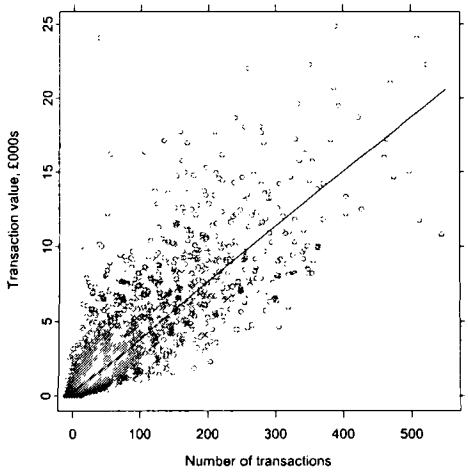


We will investigate distributions of spending behaviour in more detail in Chapter 8.

There is another problem with the analysis of credit card transactions - we only see the total amount spent on each one, and have little idea what goods the customer may have bought as part of that transaction. For example, we are aware of one customer (not in the present sample) who bought a computer in a supermarket for around £1,000. In our data, this would appear as any other transaction, since there is no flag to indicate that it was for a single item. In our design set, the ten largest supermarket transactions were for the following amounts: £299.99, £319.98, £320.29, £358.03, £371.63, £388.10, £463.00, £478.99, £525.50 and £571.52. We could speculate that the £299.99 might be for a single item, because electrical goods are often priced in this way, but what about the others? Is £478.99 also for one item? This seems rather less plausible, and of course, by chance we would expect 1 in 100 transactions to be for amounts of £xx.99.

We expect the relationship between the number of transactions and their value to be correlated, given that we selected our trade sectors to be (more or less) homogeneous groups of MCCs. We illustrate this in Figure 7.3, which shows one sector – supermarkets – it is a plot of the number and value of transactions. The line is a locally smoothed regression line, and it is almost a straight line, apart from close to the origin.

Figure 7.3 Transaction value and number, supermarkets



7.3.1 Relationships between sectors

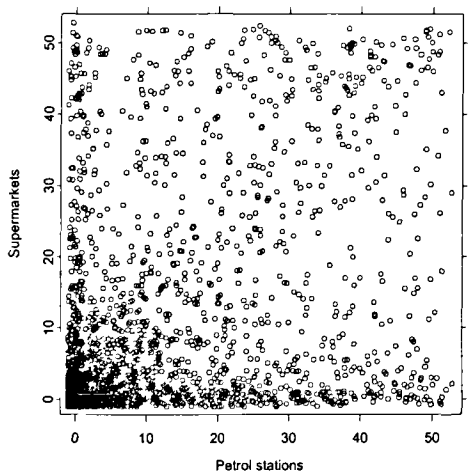
There are several ways we can represent customers' spending data in the models we construct: examples might be the amount spent on each transaction, the number of transactions or the number of sectors used. Further, each of these can be aggregated to different periods, such as annually, monthly, weekly or daily, and we described (in Chapter 3) a distortion that could arise if we used daily data. One of our aims is to be able to predict customers' behaviour, but if we seek to do that for spending, there is no obvious outcome variable that we ought to select (except possibly the total

amount spent per person in some particular time). Every sector is as suitable as any other, although we could choose to model spending in the largest sectors first, as these are most important to the business. We will therefore try to identify links that we might expect to be a feature of the data, informally in this section, and then use techniques that can assist in this exploratory data analysis.

A priori, we might expect there to be a good (in some sense yet to be defined) relationship between spending in supermarkets and petrol stations because they are both low value, frequent use, sectors. They are also what we might call 'day to day' expenditure, for which most people – we surmise – would not want to take credit, although borrowing can be distinct from spending, a phenomenon we discuss in more detail in Chapter 10. We investigate this in Figure 7.4, where we show the number of weeks that each of these sectors is used. To aid visualisation we have jittered the points along each axis, otherwise many of the points overlap, as there are only 53 discrete values for each variable.

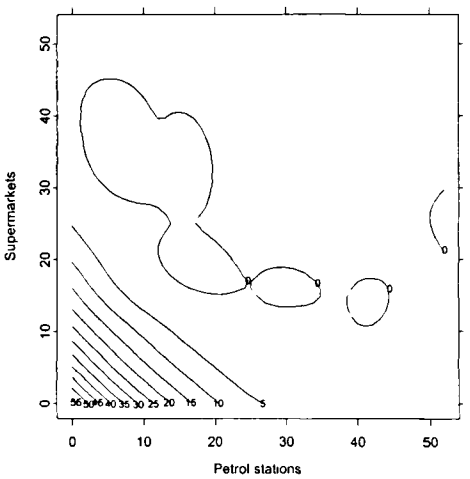
It is worth noting, though, that the correlation coefficient between these variables is 0.48. The relationship between these sectors therefore seems to be rather stronger than the simple scatter plot might indicate. If we remove those customers who spent in neither sector, the correlation coefficient drops to 0.42.

Figure 7.4 Number of weeks each sector is used



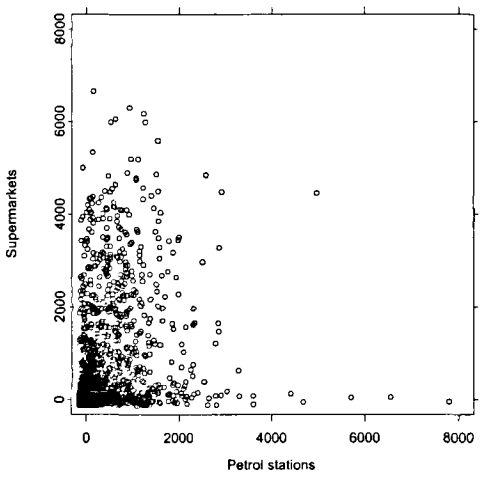
As described in Hand, Blunt, Kelly and Adams (2000), it can be difficult to use plots such as this to deduce structure, so in Figure 7.5 we use a contour plot to show that the density of customers increases as the number of weeks both sectors are used tends towards zero.

Figure 7.5 Contour plot of the number of weeks each sector is used



To the right of the line $y = 25 - x$, there are very few points: just over 19%, in fact, in 88% of the area of the plot. We show a similar scatterplot in Figure 7.6, but this time it contains the amount spent in each sector, and the correlation coefficient has fallen to 0.32. The number of transactions shows a slightly stronger relationship, with a correlation coefficient of 0.38.

Figure 7.6 Amount spent in each sector



So, although a relationship does exist between these two sectors, it is relatively weak. This suggests that any attempts to predict activity in one of these sectors is likely to be poor if we only use one other sector as a sole predictor. If we were able to find any simple links between sectors, it would be useful in a business context, because any resulting models would be simpler and easier to use than models that are more complex. The fewer variables we have to download from a mainframe, and thus to check, clean, and use, the quicker the business can act.

7.4 Frequency of use and amounts spent

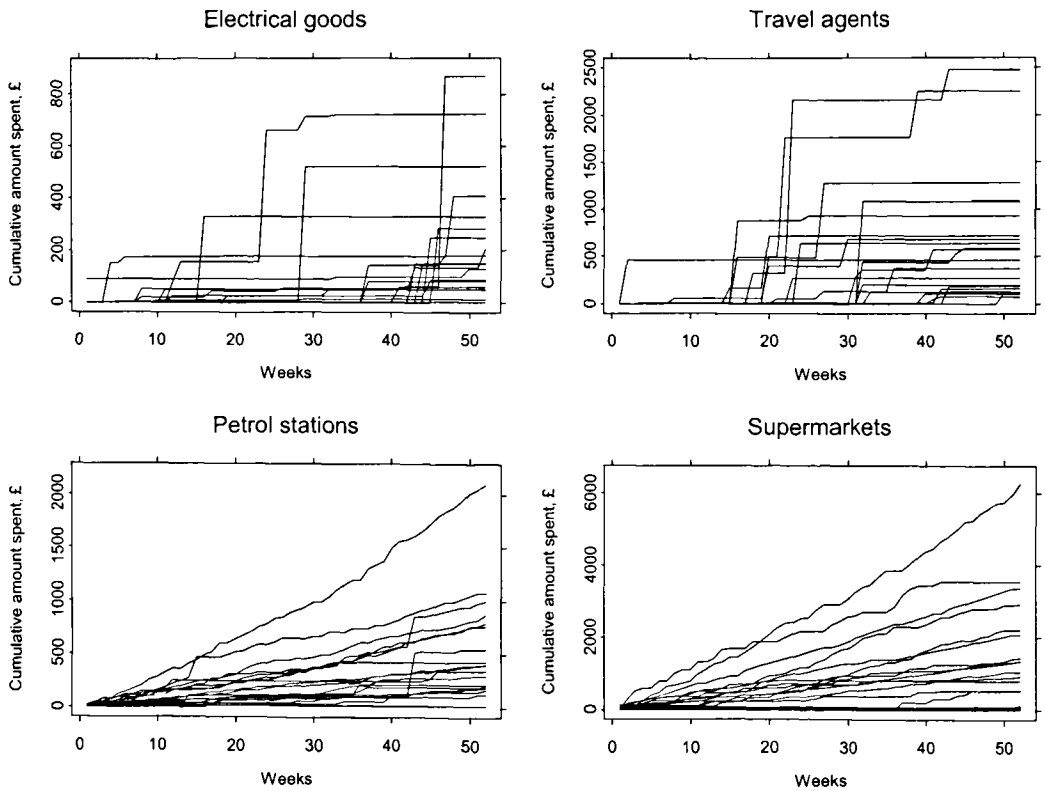
To illustrate another aspect of the differences in the way different sectors are used, we show a sample of 20 customers in Figure 7.7, and their cumulative spending week by week. The only criterion for selection was that each customer had to have used all four sectors at some point in the year. These particular sectors have been chosen to illustrate the differences between low frequency, high value (travel agents) transactions and high frequency, low value transactions (petrol stations and supermarkets). It is quite clear that, for those customers who use the latter two, they make frequent transactions in these sectors. The electrical goods retailers are somewhere between the two extremes.

Note that the plots are of weekly spending, not individual transactions. Many customers use their card more than once a week, and some make several purchases on the same day. As one example, a customer in this sub-sample (of 20 customers) made 105 transactions in supermarkets in 45 different weeks. Not only this, but she made more than one supermarket transaction on 44 days, and 33 of these apparently were Mondays. This is important, because our data set contains the date on which transactions were 'posted' to the account, and these days were (usually) neither holidays nor weekends, as we described in Chapter 3. Further, a transaction typically takes a day or two to reach a customer's account. Thus, higher spending on a Monday or Tuesday, compared to the rest of the week, is recording customers' spending at the weekend, rather than at the beginning of the week.

We have avoided some of the distortions that arise from having only 'posting' date in our data by selecting weeks as our unit of time, rather than days. If we use the latter, we would see (almost) no spending at weekends, which is obviously not the case. A

systems upgrade took place a few years after our data were extracted, and now Barclaycard records both the posting date and transaction date.

Figure 7.7 Amount spent by a sub-sample of 20 customers



Other unusual features can be seen in this plot. For example, the second highest supermarket spender (in this subset) stopped using his card in this sector in October 1996. Something seems to have happened to change his behaviour across the two years available to us (1996 and 1997). In the year to January 1996, he spent almost £6,500 on his card, and by the end of that year, his annual spend had risen to almost £12,000. By the end of 1997, it had fallen to £2,000, although he carried on using his card continuously in both years.

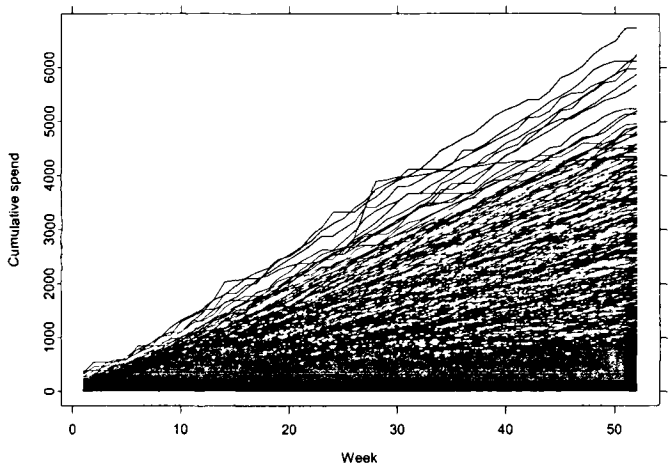
We could speculate as to why this might be the case, but at the end of 1997 he was 47, rather older – on average – than the age at which the 'usual' demographic factors, such as marriage or the arrival of children, would have caused such changes. A more likely explanation is that he took out a card with a competitor, although for obvious reasons, this information is not available to us. One consistent factor is that he did not pay interest in either of the two years, and it seems that most of his high spending months were between the end of 1995 and March 1997, and after that he only spent around £50 a month on his Barclaycard.

Another feature is that some of the curves in the petrol plots seem to have a second derivative greater than zero. We investigate this in Section 7.7, and will show that the rate of change is increasing over time, and not only that, but also that we cannot fit it with simple parametric models (for example, linear or quadratic).

However, it can prove difficult to investigate these patterns graphically with a large data set, as we show in Figure 7.8 Some of the characteristics of the sector's spending are apparent, but most are hidden in the mass of overprinting. It is also worth noting, as we did in Chapter 3, that in some instances our sample could be considered relatively small by data mining standards. If we were to apply these 'visual selection methods' to the entire data set, it could prove either time consuming, or indeed impossible, to deduce patterns among millions of consumers' behaviour. In Chapter 1 we described the work of Cox et al. (1997) who had overcome these difficulties with the development of specific software tools to detect telephone calling fraud, using methods that combined the ability of the human eye to spot patterns, together with the ability of computers to perform repetitive tasks very quickly. In the case of credit card spending behaviour, the cost justification of such

developments would be questionable. If our objective was to detect fraud, then that would be a different issue, when the potential gains could be massive. Bolton and Hand (2001) describe such an application.

Figure 7.8 Cumulative supermarket spending for all customers using the sector



7.5 Seasonality

We have only two years of data, so we will not attempt any seasonal decomposition, but will simply outline some important differences between sectors. In Table 7.3, we have chosen two sectors at random, to demonstrate how dissimilar seasonal patterns can be. Period 1 is the first four week period of the year, period 2 the second, and so on. The figures are the increase or decrease in spending compared to a straight line with no trend.

The figures confirm our prior beliefs: people spend less on 'do it yourself' in the winter, and buy more by mail order in the autumn, as Christmas approaches.

Table 7.3 Some seasonal patterns

Period	DIY	Mail order
1	-44%	-24%
2	-32%	-24%
3	-18%	-39%
4	-17%	-9%
5	39%	-16%
6	37%	20%
7	24%	28%
8	1%	-25%
9	3%	-1%
10	10%	-17%
11	-1%	73%
12	11%	46%
13	-13%	-13%

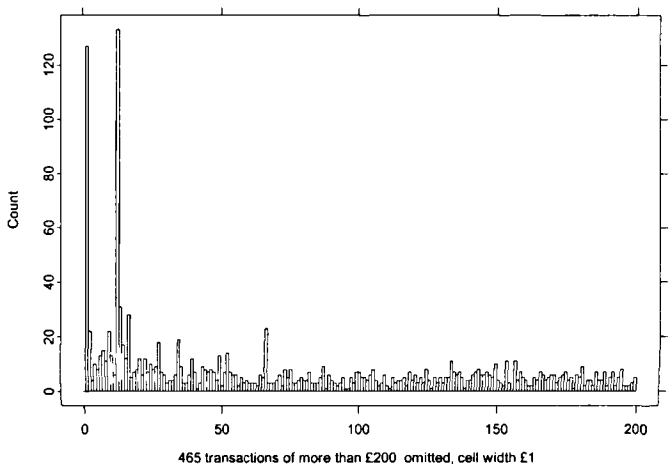
7.6 Data quality and data distortion

Our aim is to make statements about the way credit card holders behave, how we can classify them, and how we can predict future behaviour. We hope to make such statements by examining their transactions stored in a database of customer records. As Hand and Blunt (2001) note 'Transactions stored [in the credit card database] reach that position through a number of steps, and at each step there is the risk of contamination, errors, or distortions being introduced. One would like to imagine that one's credit card (and other financial) records were faultless, but no data set is free of errors, and this is especially true of large data sets and of data sets that apply to humans'. This point has been made by other authors such as Witten and Frank (2000) and Berry and Linoff (2000).

Hand and Blunt (2001) further differentiated 'between *errors*, which are simply human mistakes or machine faults of some kind, and *distortions* induced by the measurement and data collection process'. We show an example of the former in Figure 7.9, a histogram of transactions for insurance premiums (they were actually in the 'other finance' category, but many transactions in this sector are for insurance).

The two dramatic peaks at low values are due to transaction charges that were incorrectly coded. In this case the customer would not see the error; it is the way the transactions were coded for analysis. These data were extracted from an operational system designed to run customers' accounts, not for analysis of their behaviour: several consequences of such opportunistic sample extraction were described in Chapter 1.

Figure 7.9 An example of data distortion



An example of distortion in the data collection process is the built in time delay between the date at which the transaction took place and the date it enters the database, which we mentioned in Section 7.7.3. This is a distortion if one's aim is to model the date at which transactions occurred (though not, of course, if one's aim is to model the date at which they were 'posted' to the account). It is not an error or a mistake in the data.

Having identified such properties of the data, one must decide whether it is worth trying to rectify them, either for current or future data. Such rectification carries a

cost, so that cost-benefit considerations come to bear. As Hand, Blunt, Kelly and Adams (2000) point out, automatic fault rectification using computers may not be a good idea. Such exercises may well erase the interesting and unexpected structures one is seeking. Also Witten and Frank (2000) note that 'inaccurate values are hard to find, particularly without specialist domain knowledge'.

Another relevant example is given by Berry and Linoff (2000) who quoted a company executive saying: 'The data is clean because it is automatically generated – no human ever touches it'. The authors explain how 20% of transactions in the data set described files that had apparently arrived before they were sent. They continued 'not only did people never touch the data, but they didn't set the clocks on the computers either'.

Finally, as Hand and Blunt (2001) remarked 'Data mining exercises have focused attention on the poor quality of much data. It is possible that this may have the unexpected and beneficial side effect of stimulating more care in data collection – especially if it can be seen that accurate data yields accurate discoveries through data mining'.

7.7 Petrol station transactions

7.7.1 Introduction

Petrol stations are credit cards' most frequently used trade sector, and account for almost 20% of transactions, but represent a smaller proportion of value, because the average transaction size is less than in many other sectors. We will explore this sector in some detail, to see if there are any interesting patterns, either from a statistical or financial point of view.

In our sample of 3,972 customers, 2,159 people used their card in a petrol station in 1996, and made 40,068 transactions. First, we will describe the data. Our initial examination, using simple graphical tools, shows some curious and unsuspected patterns. It seems that there are two kinds of users: those who try to round their transactions to whole numbers of pounds, with certain numbers being especially favoured, and those who do not seek to produce rounded transactions. We call these two types 'rounders' and 'non-rounders' respectively.

We then develop a model for the distribution of the proportion of transactions by each customer that is rounded. This models the probability that someone seeking to produce a rounded value will succeed (and not, for example, overshoot) and that someone not especially seeking to round will do so by accident. This model fits well if we compare the two distributions by eye, but it fails a χ^2 test. We discuss why this might be, and the inadequacy of traditional tests with such large data sets.

Our next approach to this problem is to develop a model based not on people, but on their transactions, expressed in pence, modulo 100. Again, we see the phenomenon where the model fit looks good, but it again fails a χ^2 test. Our final approach is to try to classify people into two classes 'rounders' and 'non-rounders' using logistic regression. This model appears to produce a poor classification, but it is nevertheless one that could be of use - financially - to the business.

7.7.2 Petrol station spending

Figure 7.10 Petrol station transactions over time

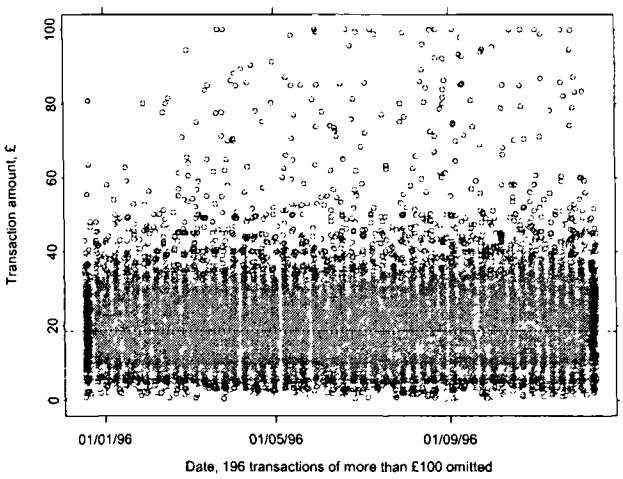


Figure 7.10 shows a scatterplot of the size of transactions in petrol stations against transaction date. The mean of all the transactions is about £20 (the median is £18.50), and the bulk of them are relatively small, with a sparse tail of higher, sometimes much higher, values. The problems of interpreting such dense scatterplots were raised earlier in this chapter, and apparently, the transaction size is not changing much over the course of the year. Notice also the vertical white stripes running down the whole plot. These correspond to weekends – transactions are not entered into the accounts on weekend days. Some white bands are thicker, and these correspond to longer breaks when Bank Holidays occurred. A better representation might be a plot of quantiles of each day's data, for instance. In Figure 7.11, we show a plot of smoothed deciles of the data we showed as a scatter plot in Figure 7.10, and now some structure is apparent. Transaction amount is flat for the first half of the year, then it increases, albeit slowly, over time.

The reason for this is quite simple, and the trend in the second half of the year can almost be eliminated if we adjust for inflation. In Figure 7.12 we show the smoothed deciles of the data when they have been so adjusted. To make this change we deflated the transaction amounts by the Office for National Statistics index 'RPI: motoring expenditure: petrol & oil' (National Statistics Online, http://www.statistics.gov.uk, 2001, series DOCU). This is imperfect for two reasons: firstly, and as we shall discuss later in this chapter, our transactions contain items other than petrol. Secondly, we described in Chapter 3 that the date we record for the transaction might not be the date it was made. Nevertheless, it provides a reasonable measure of the underlying price movements in petrol.

Figure 7.11 Petrol station transactions, smoothed deciles

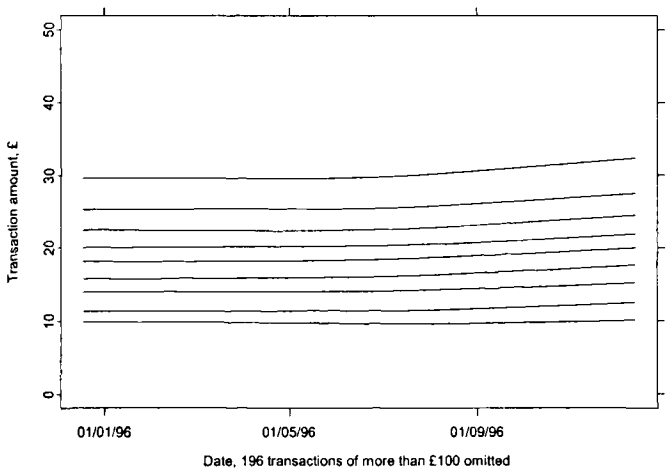
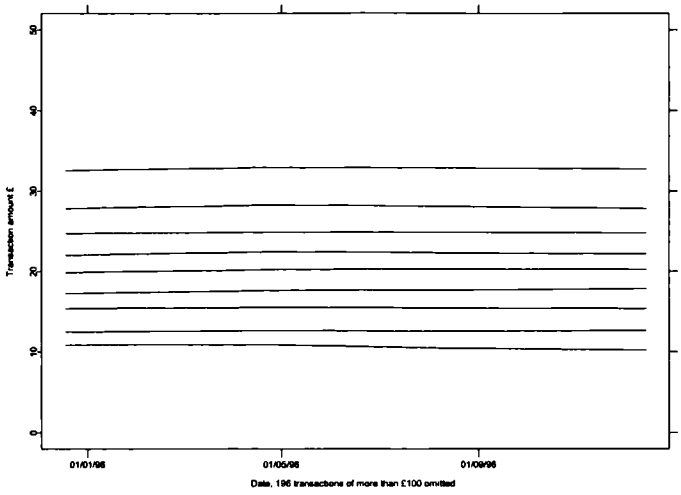


Figure 7.12 Deflated petrol station transactions, smoothed deciles



7.7.3 Intra weekly patterns

It is apparent from Figure 7.10 that there is a short term weekly pattern, but it is only the 'weekend effect'. In Table 7.4, we show the daily transaction patterns in supermarkets and petrol stations, and there is a difference between the two sectors. We derived the data in this table by looking at a year of transactions, and then summing the number of transactions by each day of the week.

Table 7.4 Proportion of transactions per week by day of the week

	Petrol stations	Supermarkets
Monday	24.2%	33.8%
Tuesday	30.4%	25.0%
Wednesday	16.1%	13.9%
Thursday	15.0%	14.4%
Friday	14.4%	12.9%

We chose supermarkets because they are the closest in frequency and transaction size to petrol stations. It is difficult to be sure why there might be difference, but it is

almost certainly because of the way in which different sectors process their transactions, and the consequent time to reach customers' statements.

For example, given the nature of supermarkets, all of this sector's transactions will be centrally processed by the acquiring bank (see Chapter 2, 'The credit card market cycle' for a description of an 'acquirer') and the supermarket. Many petrol stations, on the other hand, are individually owned or franchised, or are in smaller (often local) groups. These may then send all of their transactions overnight to a 'clearing house' which will accept all types of card transaction, such as credit, debit, petrol and charge cards, but it delays their arrival at the card holder's issuing bank.

However, petrol station transaction distributions are the same regardless of the weekday or time of year, so we conclude that the variations in Figure 7.10 are solely as a result of processing. We have not shown these analyses for the sake of brevity.

7.7.4 Conclusions – increasing value of petrol station transactions

We conclude that there is little real variation in petrol station transactions across the year, by customer, over time. There is also little variation by day of the week, and what is present is (mainly) there because of the distortions arising as a consequence of the way that Barclaycard processes transactions.

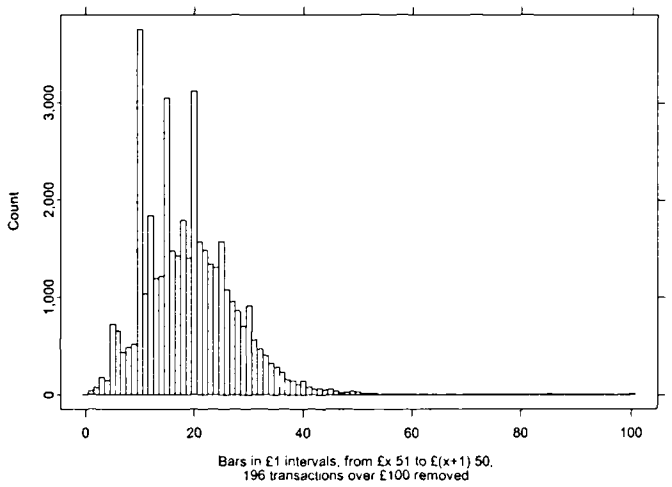
7.7.5 Individual transactions

Figure 7.13 shows the distribution of 40,068 credit card transactions made in petrol stations. The cell widths are from £x.51 to £(x+1).50 inclusive, and 196 transactions of more than £100 have been removed from the diagram. This histogram has some striking features. There is an underlying unimodal distribution that is slightly skewed to the right. When asked, beforehand, what the distribution would look like,

experienced data analysts picked on this aspect. However, the actual distribution also shows other unexpected and certainly interesting features. In particular, it shows high peaks at some values. In other contexts, such peaks sometimes arise from *digit preference*, and Hand, Blunt, Kelly and Adams (2000) describe an example of such behaviour. This is the (possibly subconscious) rounding of values to whole numbers or to numbers ending in multiples of 5 or 10. However, in the petrol consumption case, the values are recorded by machine, and digit preference cannot occur. Despite this, the peaks do occur for cells centred at values ending in 5 or 0, with a couple of extra peaks at £12 and £18. It is as if there is a substantial body of customers who deliberately seek to buy petrol costing particular amounts of money with their credit cards.

Not only this, but when we express the data in pence (modulo 100), we see that peaks occur at 5, 10, ..., 90, 95 pence as well, so it seems as if a considerable number of customers seek to round to some amount.

Figure 7.13 Distribution of petrol station transactions



The height of the peaks, relative to the rest of the distribution, was so striking that we examined some transactions from different samples of customers, in different months. This was to check that we had not inadvertently selected an unusual group of customers in the first instance. Each distribution had a similar shape, with the massive peaks at £10, £15 and £20, although at later times the £10 peak reduced in size a little, to around the same height as the £20 one, while the peaks at £25 and £30 grew.

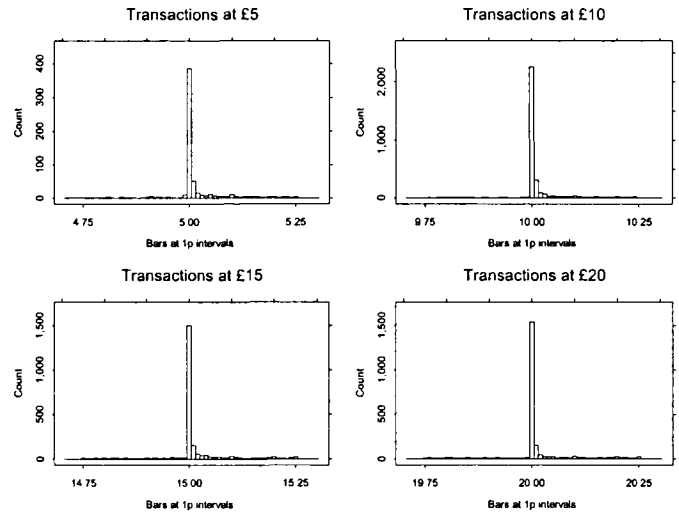
We removed 196 transactions of more than £100, and the reason should be apparent from Figure 7.13, that they are obviously different in some qualitative way from a typical petrol station transaction. In this sector, we can therefore consider them as outliers, and as there are so few of them, we will ignore them in the modelling that follows; there are another 244 which are for amounts greater than £50, and less than £100. The latter are not as distant from the majority as the still larger ones, but are still not typical of the vast majority. In most sectors, the identification of outliers is not so easy. All of the modelling that follows will therefore be carried out on a restriction of the data set to amounts of less than, or equal to, £50. By restricting our data in this way, our sample reduced to 2,120 people.

Note that petrol stations are different in one respect from most of the other sectors, which we discuss in Chapter 8. The transaction distribution is close to symmetric, with characteristic peaks, and removing data above the 99^{th} centile results in a distribution with only a minor skew compared to the other sectors.

In Figure 7.14 we present a closer examination of the distributions around some of the values producing peaks in Figure 7.13. Here the cell widths are 1p. There are clear marked peaks at the precise values, with the distribution tailing off to the right,

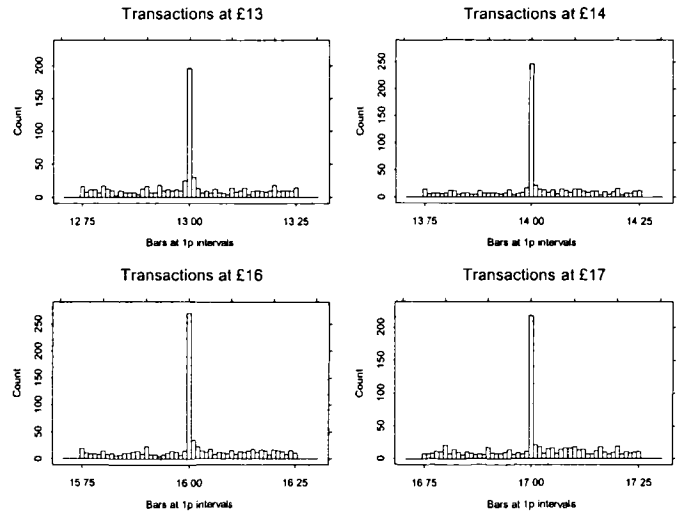
and with few values recorded below the precise values. Deliberate attempts to produce rounded values could explain this, with customers sometimes (only very occasionally, as judged by the relative heights of the histograms at exact values and at values 1p and 2p more) overshooting and hardly ever stopping too early.

Figure 7.14 Petrol station transactions at values with a peak in the overall distribution



In Figure 7.15 we show a similar set of plots around some values which do not appear as peaks in Figure 7.13. There are still central spikes here, but note that their heights are much less than in Figure 7.14. The peaks in Figure 7.14 are about 60% of the total number of transactions in the intervals shown; those in Figure 7.15 are about 30%. Also, although there might be a slight hint of more values just after the peaks than just before, this is far less striking than in Figure 7.14.

Figure 7.15: Petrol station transactions at values without a peak in the overall distribution



It is clear from all of this that some customers set out to spend rounded amounts of money, despite using a credit card. That is, that there are two distinct kinds of customer. On the one hand we have those who seek to purchase an amount of petrol corresponding to a discrete integer value of pounds, and especially values ending in 0 or 5, or purchases of value £12 or £18. On the other hand, we have those who determine the amount spent according to some other criterion (e.g. they simply fill the tank). We call the first group 'rounders' and the second group 'non-rounders'.

We can think of various explanations for adopting a rounding strategy. One is that it is a carry-over from cash, when it was convenient to pay with (for example) a £10 note. Another is that unless one intends to fill the tank, then one needs to set some sort of target and rounded values are natural targets. Other explanations for rounding are that people on a strict budget may track expenses weekly, rounded amounts are easier to remember and so may be used for checking transactions, some petrol stations require pre-payment before filling and set the pump so that it stops

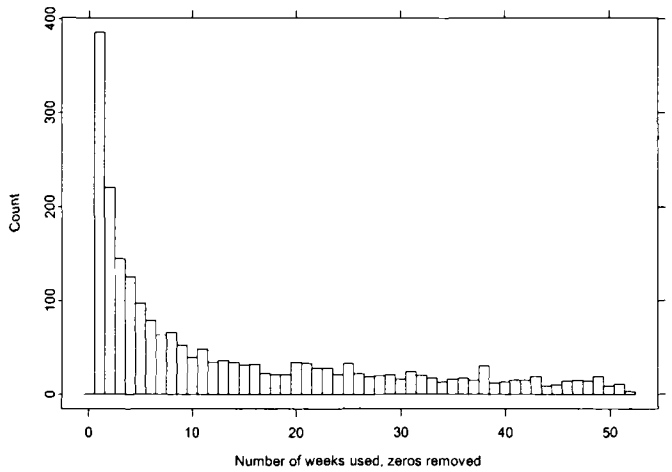
automatically at some amount (often when the forecourt is unattended, or to speed transactions).

Conversely, there are other patterns in the data too, some of which might be almost impossible to find. During the course of our investigations of the rounding phenomenon, we came across a card holder who bought petrol in amounts of £23.45 or £34.56. He deliberately targeted these values to provide him with an easy way to check that transactions on his statement are genuine.

The apparently anomalous values of £12 and £18 can be explained in terms of marketing initiatives. At the time of writing, Total, Jet and Esso petrol stations have this kind of incentive. Not only that, but the mean cash transaction in a petrol station was, according to APACS (2001c) around half the value of a credit card transaction. This is why, we believe, we see spikes at multiples of £6 as well, which are caused by the petrol retailers' forecourt incentives. These cannot, of course, easily be restricted to cash payers, because the retailer does not know, when the driver buys fuel, how she or he is going to pay.

Our data relate solely to those petrol station transactions made using credit cards and exclude any other ways of paying. It is possible that some customers normally pay cash and only occasionally use their card, for example. Indeed, as we see below in Figure 7.16, this is quite likely: some of the customers in our sample make only one credit card petrol purchase in a year. It is likely that many of these in fact buy petrol more often.

Figure 7.16 Frequency of petrol station transactions



There are other curiosities in the data. For example, there is one customer who consistently makes transactions in petrol stations of £10.35. We speculate that he rounds his petrol purchases to £10 and then also buys a newspaper. This, of course, is an illustration of a complicating feature of the data. The values of the transactions do not refer solely to petrol purchases. Around three quarters of the cost of petrol goes to the Government in tax and duty, and most petrol forecourt shops also sell a wide range of other (more profitable) items, presumably to increase their overall profit margin. Given the ready availability of these extra items, and the petrol retailing industry's desire to maximise the value from their shops, it is entirely possible, and perhaps likely, that some rounders will be influenced to buy additional items from the forecourt shop. One way of encouraging this is to site the payment counter as far away from the door as possible, so that customers have to walk past shelves full of confectionery and other items.

These items will not necessarily be in multiples of £1, so that such transactions do not appear as rounded in the data. We can see the magnitude of this rounding effect

by comparing the proportion of transactions which occur at exact multiples of £1 with the proportion we would expect if transactions were equally likely to end in any value. Of the 40,068 transactions in the data, 11,641 are rounded to multiples of £1 – that is 29% – to be compared with the 1% expected if each value was equally likely.

If we take into account the number of petrol station transactions made by each customer in our sample, then, if we assume that everyone has the probability of 0.01 of making each transaction rounded, we would expect to observe 0.149 of the people in our sample had at least one rounded transaction. In fact the proportion is 0.712. Working this backwards, and assuming everyone has the same probability of making each transaction rounded, the value of 0.712 is achieved when the probability that each transaction is rounded is 0.203. This, again, is substantially greater than 0.01. (In fact, of course, we believe that the population is heterogeneous, with some people seeking to make rounded transactions, and some not.)

Given that there are two types of people, it could be of value to discriminate between them. It may be the case that they also show different patterns of transactions in other market sectors, and perhaps advantage can be taken of this. Or, if one can identify to which group an individual belongs, perhaps a targeted marketing initiative can be cost effective (for example, attempting to induce a customer who normally spends only £10 on petrol using their card to spend more).

In attempting to discriminate between rounders and non-rounders, we must recognise that non-rounders will sometimes hit rounded values by chance (1 time out of 100 one might expect them to produce whole multiples of £1) and that rounders sometimes miss exact values (especially overshooting, as we saw in Figure 7.14).

7.7.6 Modelling the probability that a customer rounds

Define a random variable, R, as the ratio of the number of rounded petrol station transactions a customer makes (r) to the total number of petrol station transactions they make (n), where a transaction is 'rounded' if it is a discrete number of pounds in value. Let the probability that someone is a non-rounder be P, and let the probability that a non-rounder rounds by accident be p. $P(n)$ is the proportion of customers who have n transactions in the data, and P_n is the proportion of customers with n transactions who are not rounders. If we take all complete values of pounds as potential rounding targets (as Figure 7.15 suggests they are) then $p = 1/100$ to a good approximation. We will now model the data with a mixture distribution, composed of two parts, as follows.

The component of the mixture corresponding to non-rounders is a binomial distribution with parameter p, the same for all non-rounders. Of course, the number of purchases that non-rounders make is a random variable, n. However, n is observed in the data for each customer.

Assume that rounders have a probability t of rounding. It is unrealistic to suppose that this is the same for all rounders - some would be able to stop the pump less accurately than others would. Let t be a random variable, which we will assume is distributed over the population of rounders as a Beta distribution. This means that the overall probability of a rounder producing r rounded values out of n purchases is a Beta-binomial distribution. Conditional on n, the number of transactions, this leads to the following.

$$P(r|n) = P_n\binom{n}{r}p^r(1-p)^{n-r} + (1-P_n)\binom{n}{r}\frac{B(v_n+r, w_n+n-r)}{B(v_n, w_n)}$$
$$= \binom{n}{r}\left\{P_n p^r(1-p)^{n-r} + (1-P_n)\frac{B(v_n+r, w_n+n-r)}{B(v_n, w_n)}\right\}$$

where $B(v,w) = \int_0^1 x^{v-1}(1-x)^{w-1}dx$ is the Beta function, and we have a *different* model for each value of n. Each such model has only three parameters to be estimated: P, v, and w. For a given value of $n > 2$, the log-likelihood is then

$$l = \sum_i \log\left[P_n p^{r_i}(1-p)^{n-r_i} + (1-P_n)\frac{B(v_n+r_i, w_n+n-r_i)}{B(v_n, w_n)}\right] + constant$$

where, of course the subscripts n on P, v, and w arise because different models are estimated for each value of n. The restriction to $n > 2$ is necessary to avoid identifiability problems, since three parameters are being estimated for each value of n. The summation here is over the ith customer with n transactions.

To see how well this model fits the data, we compare observed and predicted histograms of R, the proportion of a customer's transactions which are rounded (again, and always below, only for those customers with $n > 2$). The predicted values are given by

$$P(R) = \sum_{n=3}^{51} P\left(\frac{r_i}{n}|n\right)P(n) = \sum_{n=3}^{51} P(r_i|n)P(n)$$

$P(n)$ here is the proportion of customers who have n transactions in the data (out of those with more than 2 transactions), and is observed directly from the data. (We are not concerned with modelling $P(n)$ here, but only with modelling rounding.) The

restriction of the summation to the range $n = 3, \ldots, 51$ is to ensure that each component model is identified. Figure 7.17 shows the observed data histogram for R and Figure 7.18 shows the predicted histogram. It seems from this that our model is capturing the main features of the process. Unfortunately, however, the match between the figures is deceptive. A customer who makes 3 transactions a year is necessarily restricted to making 0, 1, 2, or 3 rounded transactions. This means that the shape of the histogram is essentially being driven by the (observed) number of transactions that each customer makes. It seems that a different approach is needed.

Figure 7.17 Distribution of observed values of R

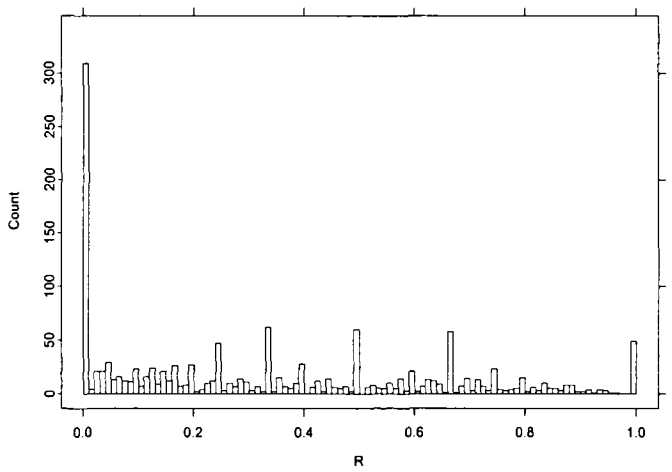
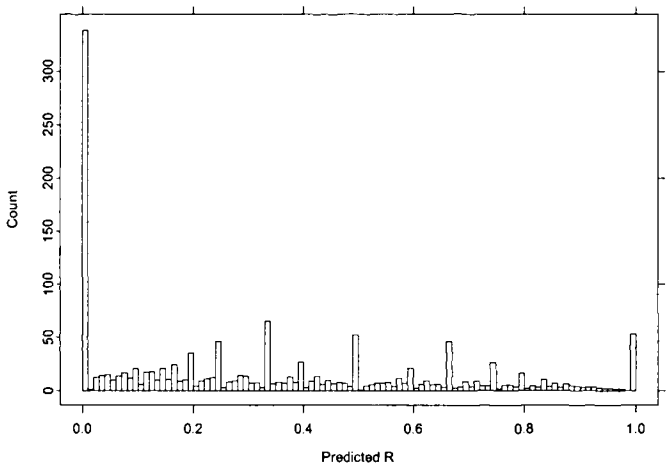


Figure 7.18 Distribution of predicted values of *R*



7.7.7 Assessing goodness of fit

Recent work in the marketing literature has tended to avoid significance tests, but has made use of some simple summary statistics to allow models to be compared. It allows rank ordering of the models under consideration, but has not addressed the degree to which one model is superior to another (Gonul and Srinivasan, 1993). Gonul and Srinivasan's comparative statistic was the Akaike information criterion (also used by Kamakura and Russell, 1989); while others have used the mean absolute deviation (Allenby et al., 1999, Allenby et al., 1998, Allenby and Ginter, 1995). With data sets smaller than the one that we have described here, Jain and Vilcassim (1991) had no success in fitting probability functions. Their objective was slightly different, but not too dissimilar – it was to describe 'interpurchase times' of two types of coffee among several hundred households, with several thousand purchase occasions.

We fitted several models to the data, using different parameterisations for v and w, and the fit of the models looked, by eye, to be good (see Figure 7.17 and Figure 7.18, for example). To assess the models we used a χ^2 test, as follows.

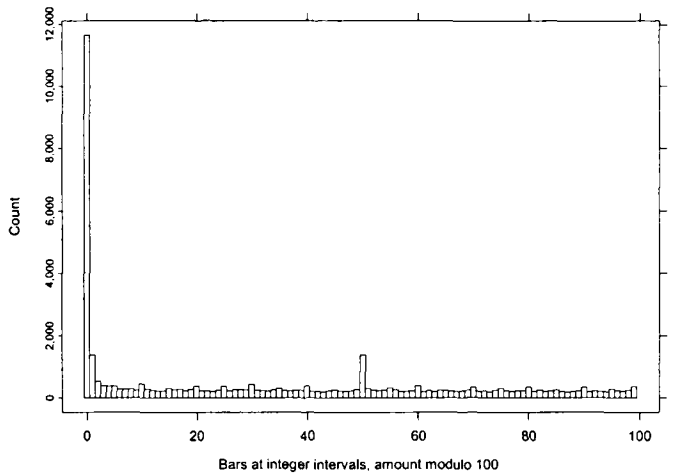
$$\chi^2 = \sum_{n=1}^{296} \sum_{r=0}^{n} \frac{\{E(r \mid n) - O(r \mid n)\}^2 O(n)}{E(r \mid n)}$$

where $E(r \mid n)$ is the number of those with n transactions whom we *predict* will have r rounded transactions, $O(r \mid n)$ is the number of those with n transactions whom we *observe* will have r rounded transactions and $O(n)$ is the number we observe to have n transactions (grouping the data so that each $E(r \mid n) \geq 5$). In each case, the predicted number was significantly different from the observed, as measured by this formulation of the χ^2 statistic. This is largely because of the large sample size, despite the closeness of the predicted values to the observed. This is the phenomenon described by Glymour et al. (1997) 'Hypotheses that are excellent approximations may be rejected in large samples; tests of linear models, for example, typically reject them in very large samples no matter how closely they seem to fit the data'.

7.7.8 Condensing the transaction size – modulo pence amounts

An alternative approach is to model transaction size modulo 100. This yields only 100 possible values: those ending in 0p, those ending in 1p, and so on, up to those ending in 99p. All of the transactions that we see are, if expressed in this form, integers. Also, because we are interested in rounding to certain values of pence, a logical way to model the data is to express each transaction in pence, modulo 100. This is shown in Figure 7.19.

Figure 7.19 Transactions in pence, modulo 100



There is a peak at 0, caused by peoples' tendency to round to whole pound amounts. Plots we have already shown demonstrated a tendency to overshoot rather than undershoot. (This was clear from the plots around £5, £10, etc, but not apparent around other values. This suggests that people do not overshoot much - it needs the larger numbers that arise with the multiples of £5 before the effect can be detected with the naked eye.) Overshooting the 'target' amount rather than undershooting is probably a feature of plastic card use in petrol stations – if transactions at 'currency note equivalent amounts' were really cash substitutes, we might expect undershooting would make more sense to the card holder. For example, if a person had only £10 in his or her wallet or purse for petrol, the implication of exceeding that amount, even by only a few pence, might be embarrassment when paying the cashier. Or, if another customer had just visited an automatic cash machine, and had a wallet containing only £20 notes, spending £xx.01 might again mean embarrassment, or a pocket full of change. With credit cards, of course, these issues do not exist, hence a

small propensity to overshoot. It is also clear from Figure 7.20 and Figure 7.21 that there are peaks at values ending in multiples of 10p and 5p.

Figure 7.20 Removing the spike at 0 so we can see the other values more clearly

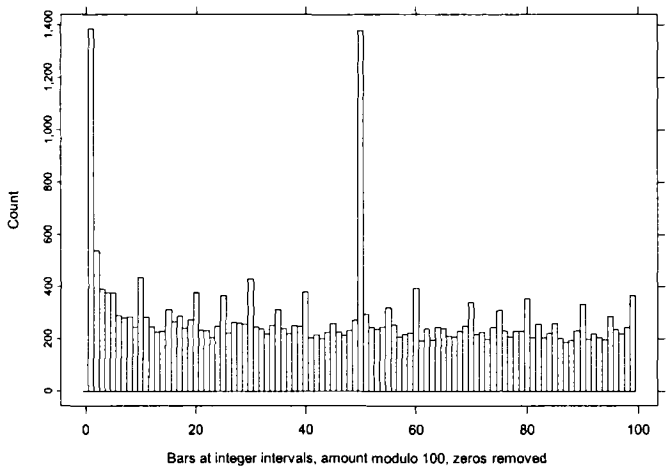
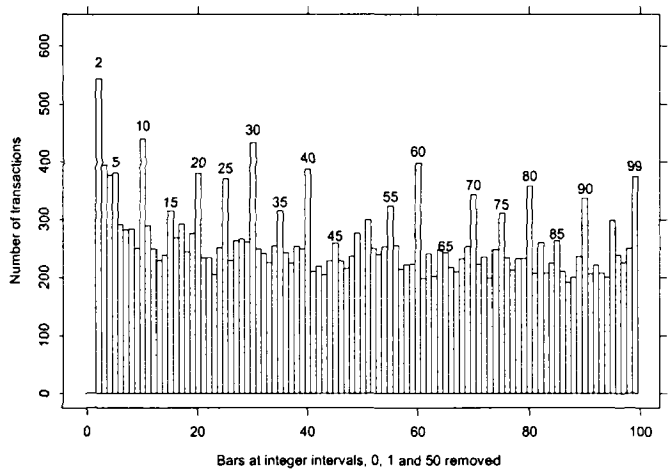


Figure 7.21 Removing the spikes at 0, 1 and 50



It is tempting to infer that customers tend to round, not only to multiples of £1, especially multiples of £5, but also to transaction values ending in 50p and also to

multiples of 5p and 10p. A more plausible explanation is the pricing policy of forecourt shops in petrol stations, where many items are priced in multiples of 5p.

7.7.9 Simple geometric model

We choose to model the grosser aspects of the distribution of rounders' behaviour by a geometric distribution, which seems reasonable, given the shape of the distribution. Non-rounders will be modelled by a uniform distribution, with $P = 0.01$, as before. Thus the probability of observing a transaction value ending (mod 100) in i is

$$f(i) = P \times 0.01 + (1 - P) \times \frac{1 - q}{1 - q^{100}} \times q^i \qquad i = 0, 1, .., 99$$

Where P is the probability that a rounded transaction is produced by a non-rounder, and q is the parameter of the geometric distribution. We are assuming a simple geometric distribution fits the mixture of rounders, although we speculate that a mixture of geometrics, with different q for each rounder, would be better - obtaining a model similar to our beta-binomial component of the earlier model.

The log-likelihood is then $l = \sum \log f(i)$, where the summation is over all transactions by all customers. Our aim is to find the values of P and q that maximise this and then to carry out a chi-squared test for goodness of fit. This test is rather easier to understand than for the beta-binomial model, because there is a predicted and an observed value for each of $i = 0, 1, .., 99$. We will then compare them using a chi-squared statistic on *100-1-2* degrees of freedom (*-1* for the known total and *-2* for the two parameters in the model).

It is obvious from Figure 7.19 that there is a spike at 50p, which presumably corresponds to the behaviour of those customers, who, having overshot the exact

pound amount by some margin, then choose the next most intuitive place to stop. This is another aspect of behaviour which we have not mentioned, and indeed, it was not apparent until we examined the modulo pence amounts. To include the spike at 50p, we have to extend the model as follows.

$$f(i) = 0.01P + \frac{1-q}{1-q^{100}}(1-P-Q)q^i + Qg(i-50) \qquad i = 0, 1, .., 99$$

To model the peaks at the other multiples of 5p we would need to include parameter for each of those as well, and we show two plots that omit the largest spikes in Figure 7.20 and Figure 7.21 where the 'local' spikes are more evident.

There are several distinct phenomena that we can see from Figure 7.21. Firstly, that there seems to be a tail up to 5 pence; secondly, the peaks at the 10s are generally higher than those at the 5s, and both of these are gradually declining, as the modulo amount approaches 99. Thirdly, stripping out both of these, and the previous exclusions too, the other values are also decaying. The models we tried all resulted in fitted values that were significantly different from the observed (using a χ^2 test), despite faithfully recording most of the distinguishing features of the data.

7.8 Logistic regression

7.8.1 Introduction

A third approach is to use logistic regression to predict whether a customer is in the 'rounder' or 'non-rounder' group.

For a rounder who makes n transactions, assume that the probability that he or she makes r rounded transactions is a binomial distribution with parameter p, B(n, p). Non-rounders are assumed to have the same distribution, but with parameter

$p = 0.01$, since this is the probability that they will (by chance) make a transaction which is rounded. The mean number of transactions that are rounded is then $0.01n$ for a non-rounder and pn for a rounder. To assign each customer to one of the classes, select a threshold t halfway between these, and assign each person to the rounder class if their proportion of rounded transactions is greater than t and to the non-rounder class if it is less than, or equal to, t. The choice of p and t is arbitrary, and more rigorous choices could be based on the individual shapes of the binomial distributions.

The logistic regressions are used to predict class membership, where the true classes are given by the comparison with t.

7.8.2 Results

The models we obtained were all very similar. Rather than describing their details, we summarise their broad characteristics:

- All predicted probabilities (of being a rounder) were in the range $0.5 - 0.7$, regardless of the value we chose for the initial probability, p, of making a rounded transaction. We used values of p between 0.01 and 0.99.
- Dividing customers into quantiles of predicted probabilities of being a rounder revealed marked differences in behaviour. For example, those we predicted to have a higher probability of being a rounder made proportionally more rounded transactions than those with a lower predicted probability. Also, the former customers made smaller, more frequent, transactions.
- These differences are enough to use for different marketing campaigns (e.g. those with the highest predicted probabilities of being rounders are more likely to make more rounded transactions).

Often, the question to be answered in a business environment is 'does this model give better predictions than were previously available' and our predictions gave markedly better results than the random selection of customers.

7.8.3 Rounders and non-rounders

We believed, a priori, that the population would be susceptible to a classification such as rounders and non-rounders, but closer examination of the data indicates that, with our current data set, it might not be possible to make a distinction between the two types of customer. Some customers appear to be rounders, as illustrated by the customer with the greatest number of petrol station transactions in our sample, of which more than two thirds were rounded. We show a sample of his transactions in Table 7.5.

Table 7.5 Number and amounts of transactions

Amount	Number	Amount	Number	Amount	Number
£10.00	8	£20.01	2	£26.00	10
£10.01	1	£20.47	1	£26.02	1
£11.00	1	£20.50	1	£26.05	1
£12.00	2	£20.82	1	£26.40	1
£13.01	1	£21.00	5	£26.80	1
£13.99	1	£21.01	1	£27.00	10
£14.11	1	£21.90	1	£27.01	1
£14.50	1	£22.00	4	£27.21	1
£15.00	9	£22.01	1	£27.98	1
£15.01	3	£22.86	1	£28.00	5
£15.49	1	£23.00	9	£28.01	2
£16.00	1	£23.01	2	£28.82	1
£16.26	1	£23.18	1	£29.00	4
£16.86	1	£24.00	5	£30.00	2
£17.02	1	£24.01	3	£30.02	1
£18.00	3	£24.02	1	£30.90	1
£18.01	1	£24.03	1	£31.00	2
£18.50	1	£24.80	1	£32.00	1
£19.00	5	£25.00	9	£33.00	2
£20.00	7	£25.50	1	£36.00	1

Many of these are explainable, such as the transactions at £20.00, and those at £20.01, where presumably the customer just failed to make a rounded transaction. However, the transaction at £20.47 would appear to be anomalous behaviour for a rounder. The patterns seen in Table 7.5 are typical of those customers with a high proportion of rounded transactions – many are at the exact amounts, while some appear to be instances of inability to stop the pump exactly at the required value. On the other hand, many values appear to have no explanation, other than the ones we described in Section 7.7, but we have no way of confirming or denying that.

Similar instances occur for many individuals – where we see a high proportion of rounded transactions, but some are for apparently unexplainable amounts. For instance, why should a customer make transactions for £14.11 and £27.21 (as shown in Table 7.5)? Attributing these to poor rounding ability seems questionable.

Further investigation is needed, and commercial insight may best be gained by a psychological investigation of customers' behaviour.

7.8.4 Potential covariates

Most of the logistic regression models gave predicted probabilities that were significantly different from the observed values. We have (in our data) a limited amount of demographic information: age (but around a third have no age recorded), sex, and a geodemographic indicator. There were few of the differences we had anticipated, and the only large difference was that customers with 'wealthier' geodemographic indicators spent more than those with 'poorer' flags. For some individuals, but by no means all, some rounding activity appears attributable to the fact that more than one person had a card on the account.

7.8.5 Conclusions about logistic regression

We discussed the use of logistic regression as a way to predict membership of the rounder and non-rounder classes. It appears that, from a practical (i.e. business) point of view, we can achieve predictions that are distinct enough to guide marketing activity. It appears that, from a practical (i.e. business) point of view, we can achieve predictions that are distinct enough to guide marketing activity, even though formal tests show significant departures from the models.

7.9 Conclusions

We detected unsuspected aspects of behaviour and went on to characterise and model it. We did this in two ways, by individual and by transaction. Some people seek to round transactions to exact values at particular £ amounts, especially £10, £15 and £20; others to whole pounds, and others to values that end in a multiple of 5p or 10p.

The extent of this was, at first, surprising. In the course of the work, we asked seminar audiences and qualitative researchers in the industry for their opinions. The existence of the structures that we uncovered (often) did not come as a surprise, and we can summarise the types of behaviour as follows.

1. The relatively infrequent petrol station user (at least on a credit card), who tends to round to a "cash note equivalent" amount.
2. Those people who choose to round to a target, of which there appear to be two "levels":
   - exact pounds (other than the ones we just mentioned), and
   - an intermediate pence amount, which is a multiple 5p or 10p.

The rest of the customers, as we speculated in Section 7.7.5, might also have a target, which is to stop when the tank is full, or to provide a convenient check on

transactions. A subset of transaction peaks in Figure 7.13 is caused by an incentive that we have already described, which is that the garage offers rewards based on spending in multiples of £6. Another explanation was put forward at a seminar: some garages offer an incentive if the transaction is rounded to an exact pound. We suspect that this might be so that garages do not have to keep a lot of change because forecourt staff do not know how a customer will pay. We also found that there were not the demographic differences that we might have expected, and we did not describe our efforts to improve the models by the use of covariates. All of our models fitted well by eye, but most failed a χ^2 test, and we described how this was largely a result of the sample sizes that we used.

However, in a business environment, the question to be answered is often not 'is our model significantly different from the observed?' but 'does it allow us to do things better than we did before completing the analysis?'. If the answer to the latter question is affirmative, then our modelling efforts have been successful. We could not give details, for commercial reasons, but described that it is possible for one incentive to save a great deal of money, of the order of hundreds of thousands of pounds.

Commercial insight may best be gained by a psychological investigation of customers' behaviour. Often, companies do this via qualitative market research, and as we mentioned in Chapter 3 (and will do so again in Chapter 10), it is one of the many approaches used by Barclaycard to gain insights about the way people use their cards.

Chapter 8

8 Characterisation of transaction patterns

8.1 Introduction – taxonomy of spending behaviour

This chapter has several parts, all with the theme of discovering characteristics of spending behaviour. There are many ways we can analyse such data – by transaction, by time, by customer and so on. First, we develop a taxonomy of spending behaviour, in several ways, but they all have similar results. One approach was first developed in 1959, then refined over subsequent decades, and it uses a univariate distribution to describe each sector. A second way is to consider transaction amounts, and how they differ by sector. In the third way, we examine the number of sectors used.

After these investigations, we describe graphical models used as a simple exploratory technique, to try to find links between different sectors. Following this we describe the use of conditional probabilities and odds ratios, also and our aim is to find connections between sectors, and compare and contrast these with association rules, a technique from computer science. Next, we show, using principal components analysis and cluster analysis, an example of a phenomenon we described in Chapter 1. 'Data mining generally deals with messy, distorted data, possibly from samples that have not been properly constructed, and if the formal inferential tools of statistics are applied, their results may need to be viewed with some caution'. In this case, we demonstrate that each of these techniques has its limitations, and reveals little structure in our data.

Finally, we show how the conditional probability of sector use links to the first analyses, and that this allows us to classify sectors into a taxonomy of spending.

8.2 Number of transactions

In this section, we describe our attempts to develop a taxonomy of spending distributions. We will describe three possible ways of doing this: the first based on the number of transactions made by individual customers, the second on transaction amounts in each sector, and the third is based on the number of sectors used, again by individual customers.

8.2.1 The negative binomial distribution

We described, in Chapter 1, how Ehrenberg (1959), Chatfield et al. (1966), Goodhardt et al. (1984) and Chatfield (1986) used the NBD to model consumer purchasing behaviour. In this chapter, we will describe our application of this technique to credit card transactions. First, we will describe use of the NBD to fit four sectors at the extremes of use (as measured by low and high frequency of spending) – travel agents, furniture shops, department stores and petrol stations. Use of the NBD requires stationary purchasing behaviour, a reasonable approximation because we are only considering a year of data, and remembering the cumulative plots we saw in Chapter 7. The only sector that was non-stationary was petrol stations, but that was mainly because of the effect of inflation, which does not affect (at least directly) the number, as distinct from value, of transactions.

Glymour et al. (1997) noted that one of the problems often seen with 'available data' – which we referred to in Chapter 1 as 'opportunistic' data – is, among others, that 'frequency distributions of samples may not be well approximated by the most

familiar families of probability distributions'. However, in our case, as for the authors mentioned in the last paragraph, use of a relatively simple distribution – the NBD – produces good fits to the data. 'Good' in this case means that, using a χ^2 test, the predicted values are not significantly different from the observed values. Further, although we will show that more than half of our sectors do not meet this criterion, many only diverge from the NBD at high quantiles of data, typically at around the 0.98 to 0.99 centiles. This is more striking than at might first appear, because we saw in Chapter 7 that the large sample size was a contributory factor to the poor fits we saw then. Often there are one or two apparently anomalous values which contribute a great deal to a high χ^2 statistic, but, as we show later, the NBD provides a very good fit.

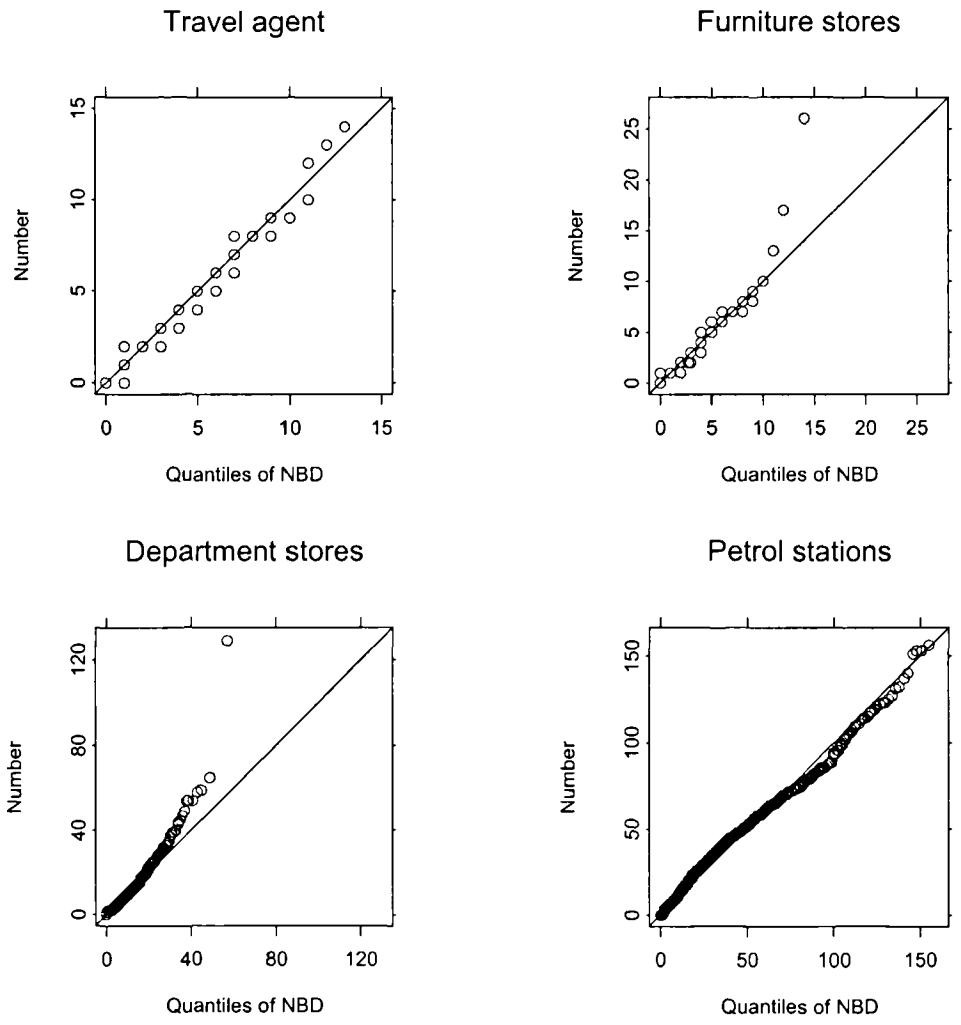
Let the number of transactions made by each customer over the year of our data be independent and follow a Poisson distribution with a constant mean. Further, let average purchasing rates of individuals differ according to a Gamma distribution across the whole sample. Thus, the frequency distribution of purchases should follow an NBD, which may be written as follows.

$$P(r) = \frac{\Gamma(k+r)}{\Gamma(k)r!}\left(1+\frac{m}{k}\right)^{-k}\left(\frac{m}{m+k}\right)^{r}$$

Let $r = 0, 1, \ldots,$ be the number of purchases made by each customer, where $m > 0$ is the mean of the distribution and $k > 0$, often called the exponent. The models we fit can then be expressed as *NB(k, m)*. Fitting this distribution to selected sectors results in the quantile – quantile plots shown in Figure 8.1.

Fitting the NBD to all sectors resulted in some fits that did not show significant differences between the observed and predicted values, while others *were* significantly different. The fit worsened in these sectors at high numbers of transactions, which is similar to the phenomenon observed by Chatfield et al. (1966) – that there are occasional 'excessively heavy buyers'. We discuss this divergence from an adequate fit in Section 8.2.4, where we argue that for many practical purposes, the models are good enough to be usable in a business context. Also, notice that the more frequently a sector is used (of the four we have shown these are department stores and petrol stations), the worse the fit becomes.

Figure 8.1 Quantile – quantile plots for four sectors



8.2.2 Numeric predictions for two of these sectors

Table 8.1 shows observed and predicted values for one of the sectors, travel agents, and in Figure 8.2 we show department stores. We could have shown any sectors: the observed and predicted values are similarly close in each case.

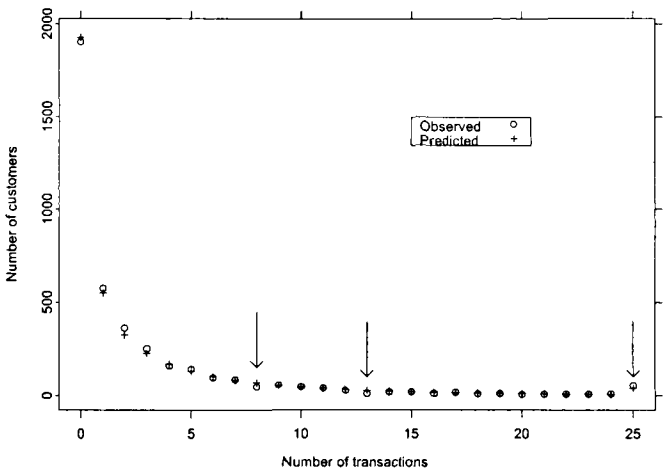
Table 8.1 Travel agent transactions

Number of transactions	Frequencies Observed	Predicted
0	2,757	2,749.6
1	590	634.1
2	316	275.8
3	141	138.7
4	76	74.5
5	37	41.5
6	24	23.7
7	10	13.8
8	11	8.1
9	6	4.8
10	1	2.9
11	0	1.7
12	1	1.0
13	1	1.0
14+	1	0.8

A χ^2 test to compare observed and predicted values has χ^2 = 12.0, which, at 7 degrees of freedom, results in a significance probability (SP) of 0.102 (the small number of degrees of freedom is caused because we pooled several of the smaller categories to ensure that all of the predicted frequency counts were at least 5). So, we have no evidence in favour of rejecting the null hypothesis, and conclude that the data are adequately fitted by a negative binomial distribution. We note, almost parenthetically, that Ehrenberg (1959), Chatfield et al. (1966) and Chatfield (1986) set the expected proportion of zeros to its observed value, but we have imposed no such restriction, and still have a good fit.

Next, the observed and predicted values for department store transactions, this time as a chart, in Figure 8.2, to show how close the fitted values are to the observed. We can see the good visual fit, as we did for a different (but related) data set in Chapter 4.

Figure 8.2 Department store transactions



Using a χ^2 test to compare observed and predicted values, we have χ^2 = 47.2, 23 degrees of freedom, and an SP of 0.002. Thus we reject the null hypothesis, and conclude that a negative binomial model is not adequate for modelling the number of department store transactions. However, as we noted in Section 8.2.1, the model fits well over most of the distribution, with a poor fit at a high number of transactions. There is another feature causing the χ^2 statistic to be high, and it is the low numbers of observed transactions at $n = 8$, and $n = 13$, compared to the predicted values. These three points are indicated with arrows in Figure 8.2. It happens in several other sectors too, and it is predominantly at these points where the χ^2 statistic

increases markedly, as well as the tail at high numbers of transactions, which is usually under predicted.

8.2.3 All sectors

We show all sectors in Table 8.3, and the centiles at which the observed values differ from the predicted values. For those sectors in which the predicted values are significantly different from the observed, the fit mainly diverges at high quantiles, typically the 98th or 99th centile. Or, like department stores, there may be an anomalous value in the body of the distribution. The ticks show the sectors where there was no significant difference between the predicted and observed values. The three letter identifiers are as we showed them in Chapter 7, but we repeat them in Table 8.2. Petrol stations are a little different from most other sectors, because the fit is very good up to the 76th centile, after which it worsens a little, but only becomes very poor after the 98th centile.

Table 8.2 Abbreviations used for each sector

ban	Bank (cash)	gar	Garage	smk	Supermarket
che	Chemist	hot	Hotel	tra	Transport
cin	Cinema	jew	Jeweller	trv	Travel agent
clo	Clothing shop	mai	Mail order	ofo	Other food
ctn	CTN	off	Off licence	osh	Other shop
dep	Dept. store	pet	Petrol station	ole	Other leisure
diy	DIY	pub	Public houses	ofi	Other finance
ele	Electrical	res	Restaurant	oth	Other
fur	Furniture	sho	Shoe shop		

Table 8.3 Centiles at which observed values diverged from predicted values

ban	✓	gar	✓	smk	0.90
che	0.98	hot	0.99	tra	0.99
cin	✓	jew	✓	trv	✓
clo	0.98	mai	0.99	ofo	0.99
ctn	0.99	off	0.98	osh	0.96
dep	0.98	pet	0.76 / 0.98	ole	✓
diy	✓	pub	0.99	ofi	✓
ele	0.98	res	0.97	oth	0.96
fur	✓	sho	✓		

For the vast majority of the data, as we would expect from Table 8.3, the NBD fits the observed values well, and in Table 8.4 we show the parameters for each of the sectors.

Table 8.4 Parameters for each of the sectors' NBDs

ban	*NB*(0.12, 2.00)	gar	*NB*(0.29, 0.94)	smk	*NB*(0.18, 9.16)
che	*NB*(0.18, 1.33)	hot	*NB*(0.20, 1.11)	tra	*NB*(0.17, 1.08)
cin	*NB*(0.20, 0.46)	jew	*NB*(0.38, 0.29)	trv	*NB*(0.36, 0.64)
clo	*NB*(0.38, 2.40)	mai	*NB*(0.19, 1.09)	ofo	*NB*(0.05, 0.27)
ctn	*NB*(0.07, 0.14)	off	*NB*(0.08, 0.56)	osh	*NB*(0.51, 6.57)
dep	*NB*(0.32, 2.74)	pet	*NB*(0.20, 9.77)	ole	*NB*(0.20, 0.45)
diy	*NB*(0.27, 2.77)	pub	*NB*(0.05, 0.06)	ofi	*NB*(0.28, 0.48)
ele	*NB*(0.43, 0.68)	res	*NB*(0.22, 2.32)	oth	*NB*(0.48, 3.70)
fur	*NB*(0.23, 0.43)	sho	*NB*(0.42, 0.73)		

8.2.4 Conclusions

The number of transactions per sector is adequately modelled by a negative binomial distribution for certain sectors and typically these are the ones with small numbers of transactions per year. Often, the distributions fit well across the vast majority of the data. Also, apart from petrol stations, the models which were significantly different diverged where the observed points were above the line of perfect fit, which means that, in general, there were too many large values to be fitted by an NBD. Again, it

is similar to the phenomenon mentioned in Chatfield et al. (1966): that of 'excessively heavy buyers', and in Goodhardt et al. (1984), who mention the wide applicability of the NBD across a broad range of product fields. In our case, the interpretation would be across a range of sectors. We seem to be seeing a phenomenon in which there are a small number of heavy users in many sectors, and this causes the fit to worsen at high values. This suggests a mixture model may be appropriate, consisting of the NBD and a model for heavy purchasers.

There could be two approaches to heavy purchasers, especially as high spenders tend to generate high value for the company. We would seek, in future work, to investigate such customers more fully, although that would necessitate much larger samples (than is the case with our data), or that we would need to sample high spenders disproportionately. This is a suitable topic for future work, given the likely value to the business of high spenders.

The NBD is thus a useful distribution to use to describe sector spending, and with only two parameters to estimate (the mean and exponent); it is easy to use. Such uses would be the easy calculation of sector 'norms' against which we could assess particular sub-groups of customers for whom we were to develop incentive schemes. We would then be able to compare the modelled fit of these targeted groups with the overall distribution to measure effectiveness of our promotions.

8.3 Transaction amounts by sector

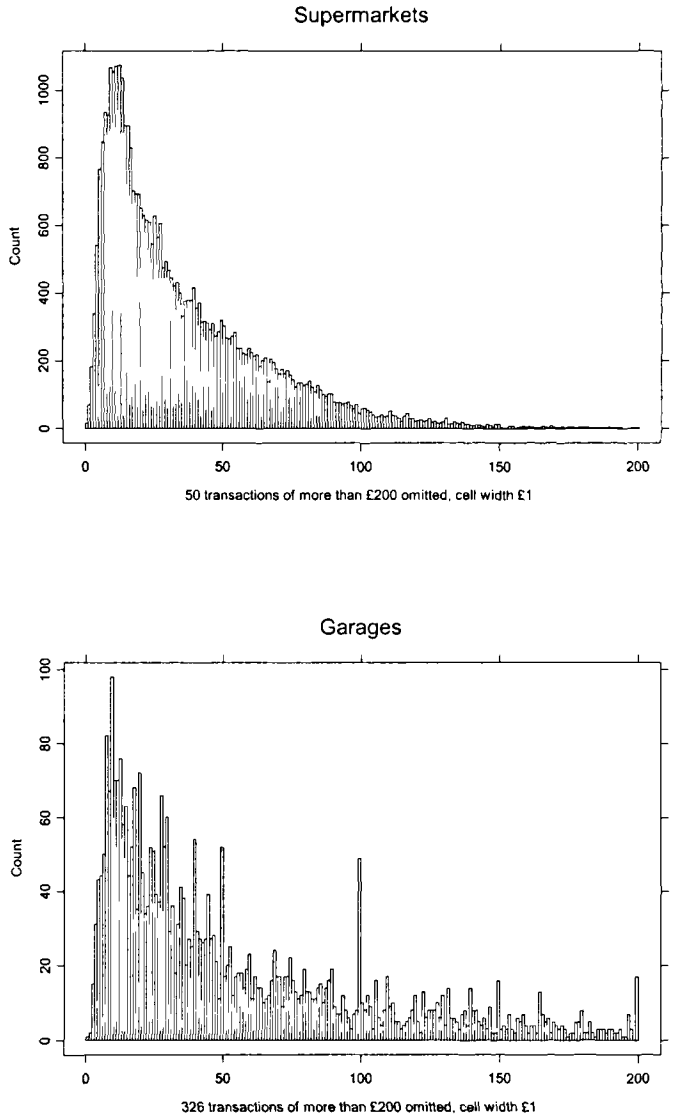
Here we consider individual transactions, and their characteristics, rather than the people making them. There appear to be three broad classes, but they are driven largely by the pricing policies of merchants in those sectors. We can characterise

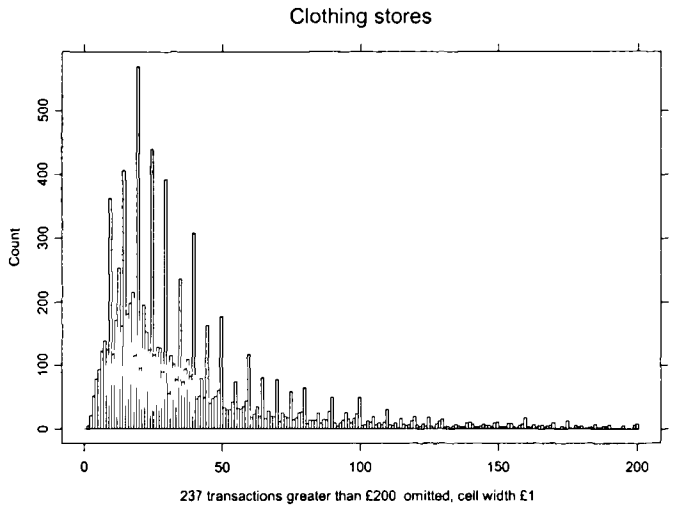
them according to the number of distributions it might take to fit the data adequately. We give three examples below (Figure 8.3) for supermarkets, garages and clothing stores. The first of these could be modelled by a single distribution, and the second by a mixture distribution, but only comprising two or three extra components for the occasional 'anomalous' spike. The final type could be modelled by a mixture of many similar right skewed distributions, with varying degrees of discreteness in their values.

The petrol station distribution we described in Chapter 7 requires a different approach.

Note that we have truncated all three histograms to £200, but have indicated how many transactions that action omitted. All of the cell widths are £1, and they have to be set to a value that is not too broad, otherwise the characteristic spikes disappear. Further, sectors such as clothing and department stores have many transactions of £xx.99, £yy.98 and £zz.97, which is a consequence of single and multiple items being bought as part of the same transaction, all for 1p less than an integer number of pounds. We will not pursue the characteristics of these distributions any further, because we cannot influence them directly, as their shapes are caused by pricing policies in these outlets. This contrasts with petrol stations. If any of our initiatives were to result in customers making more transactions, the locations of the distributions would be likely to change, which could provide one method for assessing incentives.

Figure 8.3 The three types of spending distribution





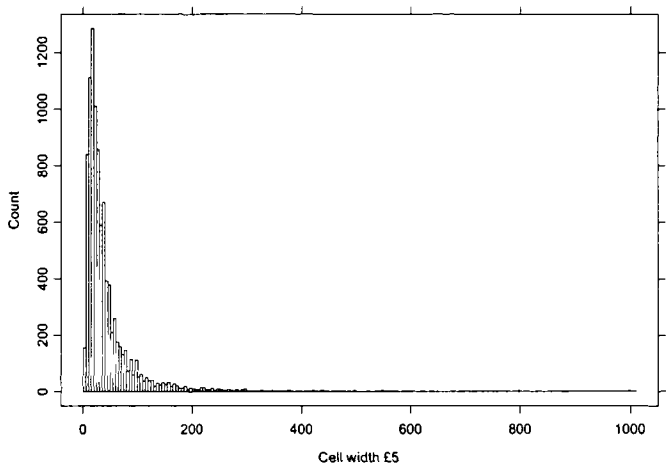
The spike at £100 in garages is striking, but further examination of the individual Merchant Category Codes in that sector provides some degree of explanation. This is purely supposition, because, as we described in Chapter 7, we only see the amount of each purchase and where it took place: we do not know what the customer bought in that transaction.

Of the 49 transactions between £99.01 and £100, thirty three were for exactly £100, and six were for £99.99. All of the latter were made in 'automotive parts and accessories stores', and are likely to be for single items; 20 of the former were made in 'automotive service shops (non-dealer)', and could be some sort of fixed servicing cost. Transactions for £100 could be the deposit on a vehicle, in which case we might not expect to see a purchase for the remaining amount, which would often be in the form of a cheque, many dealers being unwilling to accept credit cards for thousands of pounds.

In Chapter 1 we noted that data need to be analysed with a degree of sensitivity to the task in hand. To illustrate this, consider Figure 8.4, which shows the same data as the last one in Figure 8.3; the only difference is that we have now allowed the data to

shape the histogram. There is the same number of cells, but each one now has a width of £5, and the characteristic structure has all but disappeared.

Figure 8.4 Clothing stores, all transactions



The twenty six sectors could be grouped logically, as follows, according to the number of elements necessary to fit a suitable mixture distribution.

- supermarkets, mail order, 'other', DIY, restaurants, cinema, chemists, 'other' food
- garages, furniture stores, hotels, travel agents, transport, off licences, 'other' shops, 'other' leisure, 'other' finance, CTN (confectioner, tobacconists and newsagents), public houses etc.
- clothing shops, shoe shops, jewellers, department stores, electrical goods retailers, cash
- petrol stations.

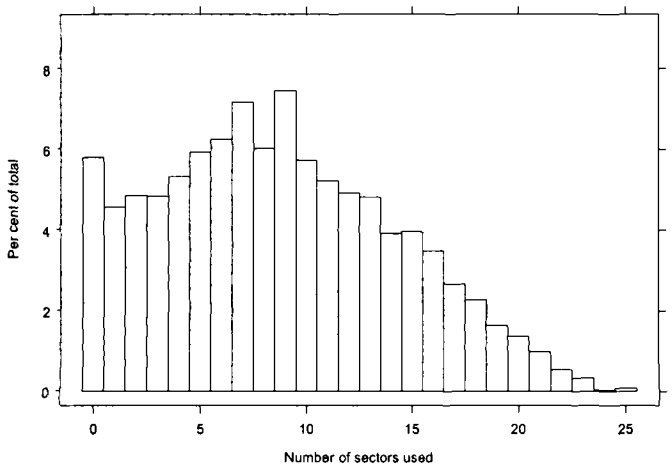
8.4 Number of sectors used, by sector

8.4.1 Introduction

In Figure 8.5 we show the distribution of the number of sectors used per customer. There is a peak at zero (customers who have a card who but have not used it in the

last year), a maximum at 9, after which it falls steadily (almost linearly in fact) to the maximum number of sectors used, 25. Each customer appears only once in this histogram, and we will examine histograms conditional on use in each sector in Section 8.4.3. This is to investigate the differences we suspect exist.

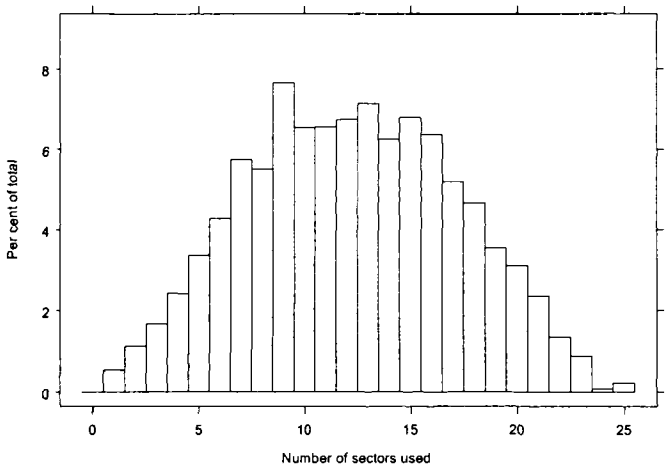
Figure 8.5 Number of sectors used – all sectors



8.4.2 'People sector occasions'

If we produce a histogram of the number of sectors used that is conditional on use in any particular sector, it will necessarily be different from Figure 8.5. No conditional histogram can contain non-spenders, because each customer in the distribution must have used – at least – the sector shown. We can make the comparison between the 'all sectors' plot in Figure 8.5, and those conditioned on an individual sector in direct in the following way, and we show the results of that in Figure 8.6.

Figure 8.6 'People sector occasions'



Let **X** be a 3,972 × 26 element matrix where, for each customer i in sector j, we have the following

$$x_{i,j} = \begin{cases} 1 \text{ if spend in sector } j \geq 0 \\ 0 \text{ if spend in sector } j = 0 \end{cases}$$

Now, let **y** be a 3,972 element vector where

$$y_i = \sum_{j=1}^{26} x_{i,j}$$

Thus, **y** is a vector of the number of sectors used by each customer. Now, form the vector **z**, where for each customer i, who has used y_i sectors,

$$z_i = \text{rep}(y_i)$$

where rep() is the repetition of each y_i, y_i times.

Thus **z** is the vector of 'sector occasions', and in Figure 8.6 each customer i appears y_i times, and it is the distribution of **z**. Thus, the people with transactions in only one

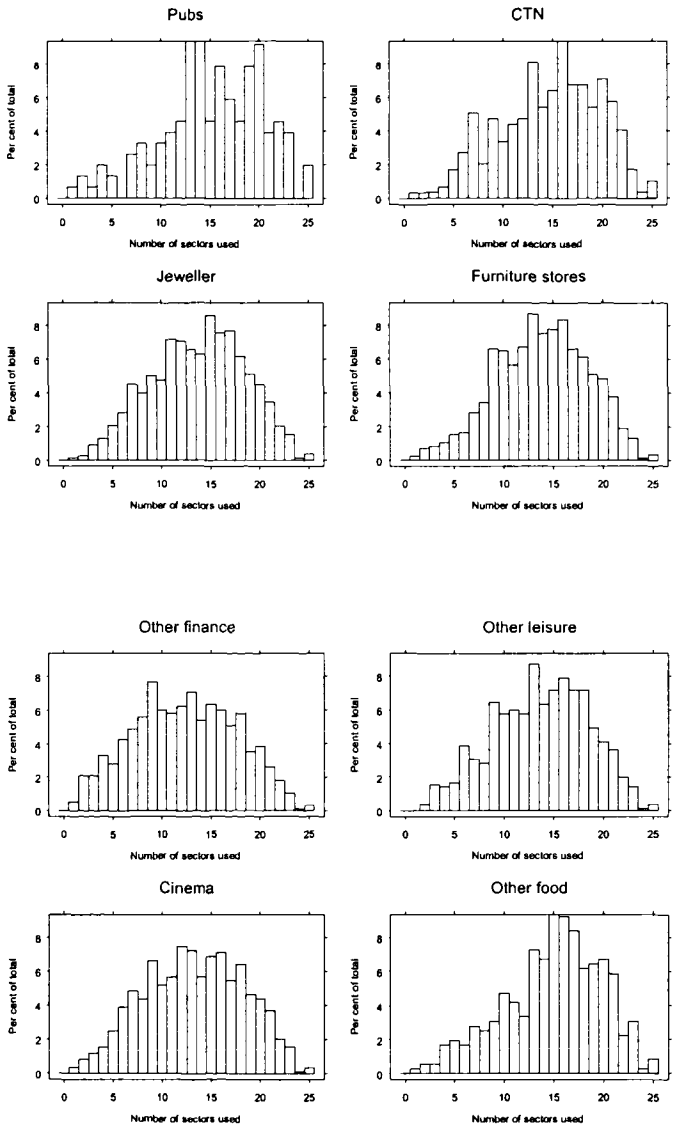
sector appear only once, while those using n sectors appear n times. There are 34,740 observations in Figure 8.6. This is directly comparable to the distributions conditioned on sector use. In each of the latter, we are measuring 'sector occasions' conditional on use in the named sector, while in Figure 8.6 we are measuring total 'sector occasions'.

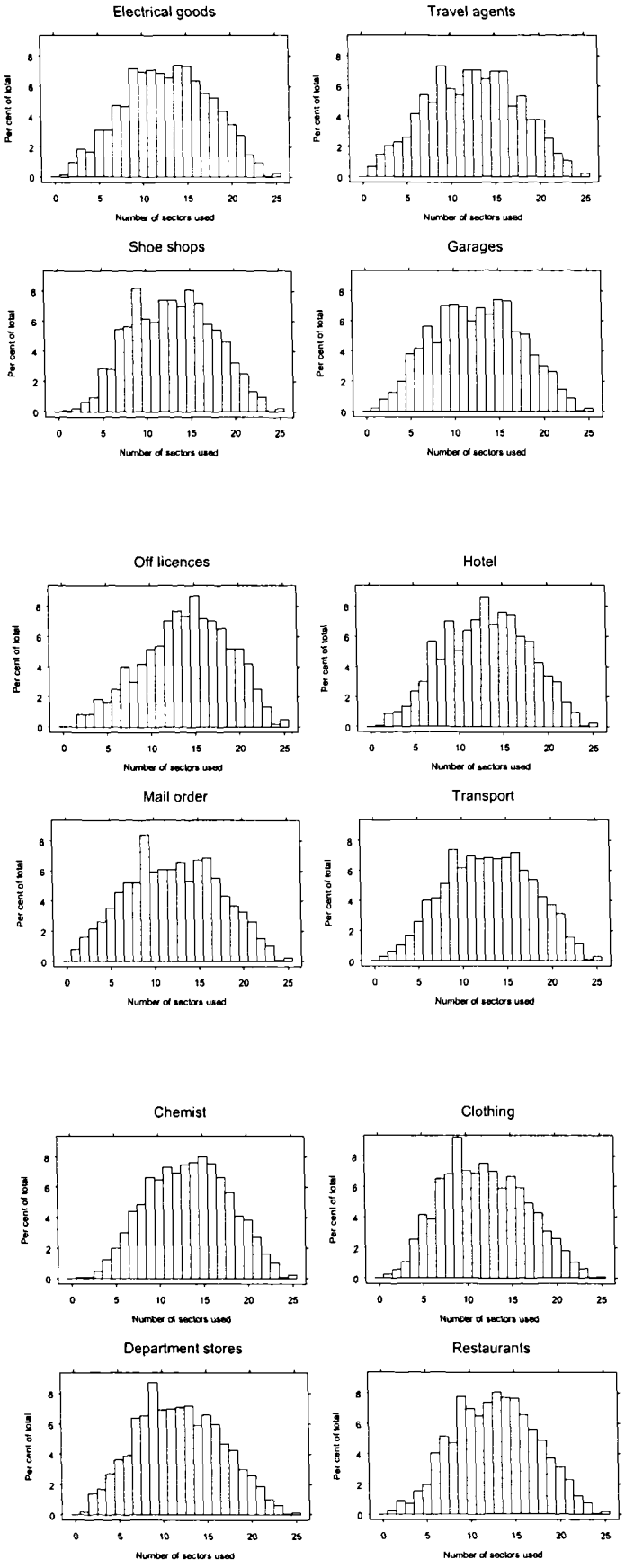
8.4.3 Sector use by all other sectors

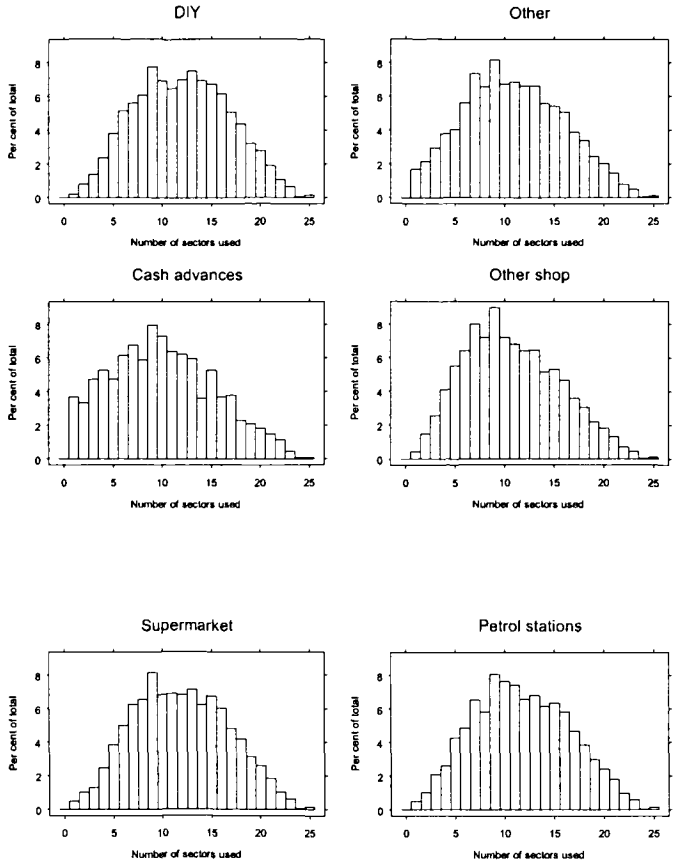
Each plot in Figure 8.7 shows the number of sectors used by customers conditional on the sector named in each individual plot, and some of the distributions are quite distinct. They are all plotted to the same scales, both horizontally and vertically, to aid easy comparison, and are arranged in order of the variance of the predicted NBD for each sector's of transactions, lowest to highest. We chose to rank them in this way because we needed display them in some order, and, given the goodness of fit of the number of transactions to the NBD, this seemed logical. Randomly sorted would have been possible, but if there was any structure to be found in these data, it was consistent to use some structure that we had already discovered. Note the relationship that is apparent – the smaller the variance, the more left skewed is the distribution, and vice versa.

Most customers will appear in more than one of the plots, because of the way we have drawn them. In each case, we simply looked at people spending in the sector we have shown, and then plotted the number of sectors they used. Thus, for example, those customers who have $n = 10$ will appear in 10 separate histograms. Compare this with Figure 8.6, where each use of each sector appears once; Figure 8.7 has (in each plot) use of each sector, but every one is conditional on use in the named sector.

Figure 8.7 Number of sectors used, by sector




Electrical goods
Travel agents
Shoe shops
Garages
Off licences
Hotel
Mail order
Transport
Chemist
Clothing
Department stores
Restaurants
Per cent of total
Number of sectors used



Judging by the shape of these distributions, we seem to be seeing a phenomenon in which a customer who is happy to use infrequently used sectors is prepared to spend in most others as well. This is the first instance of the notion of 'propensity to spend', which we will describe more fully in Section 8.7.

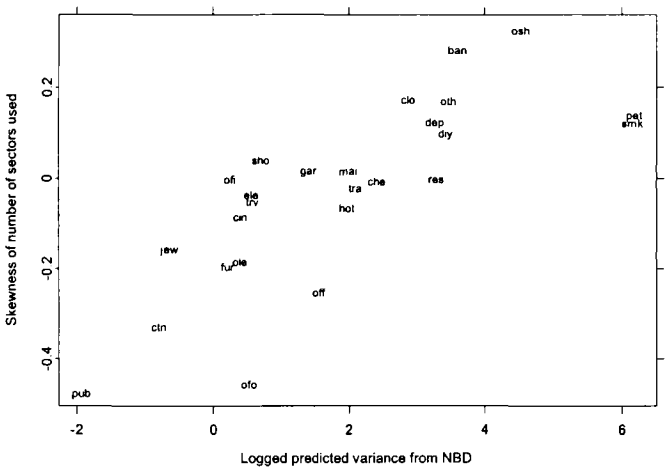
8.4.4 Number of sectors and transaction amounts

The number of sectors used can be linked to transaction amounts as follows.

We noted that the skew tends to increase from negative to positive as the predicted variance (from the NBD) for each sector increases. We show these two results in Figure 8.8, where we show the coefficient of skewness for each distribution, compared to the predicted variance of each sector, as fitted by the NBD. We have logged the variance, because petrol stations and supermarkets had much higher

values than most others. Note the order in which the sectors appear, because a similar one will occur in Section 8.6, conditional probabilities.

Figure 8.8 Skewness of number of sectors used and the NBD



8.4.5 Conclusions about a taxonomy of spending behaviour

We discuss this further in Section 8.9, but already there are several options for forming a taxonomy of use, as follows.

Sectors in which transactions are rounded by choice (e.g. petrol stations), sectors where transactions are rounded by pricing policies (e.g. department stores), sectors where there is no rounding (e.g. supermarkets) and sectors where there are a small number of anomalous spikes (e.g. travel agents). Each of these fits into the mixture distribution framework we described in the Section 8.3.

Another conclusion is that different sectors have different transaction size patterns. This is obvious in retrospect, but not necessarily beforehand, and some were definitely not all obvious in retrospect – petrol station transactions being a good example of this. There are surprising patterns in the transaction size distributions,

with the petrol histograms illustrating this. There might be other interesting patterns that we have not yet discovered, or that might be hidden by our grouping of Merchant Category Codes into 26 trade sectors.

However, no taxonomy is likely to provide a discrete set of categories, but may allow us to measure position along a continuum of various types of use. We can choose to segment these continua according to the task in hand – we discuss the difference between a *segmentation* of convenience and a statistical *clustering* in Chapter 10.

8.5 Graphical models

All of the analyses we have described so far in this chapter have been univariate. We will have a much more usable (from a marketing point of view) set of results if we are able to link activity in different sectors, and deduce how it is related. Graphical models could provide a useful way of investigating the dependence of all sectors simultaneously. As Whittaker (1990) says, graphical modelling 'provides useful information about the relative importance of the interactions'. This is important when we are performing exploratory analysis of spending data, because, and as discussed in Chapter 7, we do not have an obvious response variable, or set of predictor variables. We could predict total spending, of course, but that is simply a sum of the spending in individual sectors.

Graphical models are thus appealing, at least to try to do form an initial view of relationships between variables. We have already described several ways of representing transaction activity, and others might be as follows.

- Total amount spent in the sector (this could be over any arbitrary period, although we will concentrate on a year).

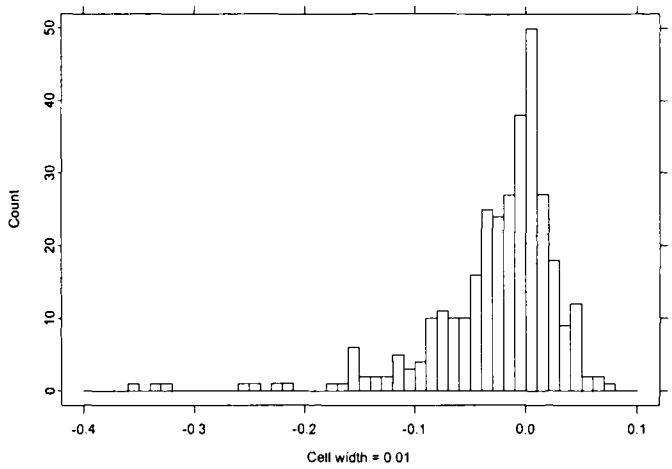
- Number of weeks that customers used each sector.
- Proportion of each customer's spending that takes place in each sector.

Note that none of these variables are normally distributed, most are not even approximately so, but the technique might nevertheless prove useful to illustrate some relationships. For the sake of brevity, we will only describe models built using the total value of transactions in each sector, because the alternatives resulted in similar models. Transforming the variables does not improve the situation, but introduces a new set of distortions – for example, many of our variables were right skewed, but a log transformation resulted in a left skewed distribution.

Following the methods outlined by Whittaker (1990) and Edwards (1995) we constructed the inverse correlation matrix for spending in the 26 trade sectors, and scaled it to have unit entries on the main diagonal. We then examined this matrix, with the intention of leaving only those edges in the model that had values markedly different from zero.

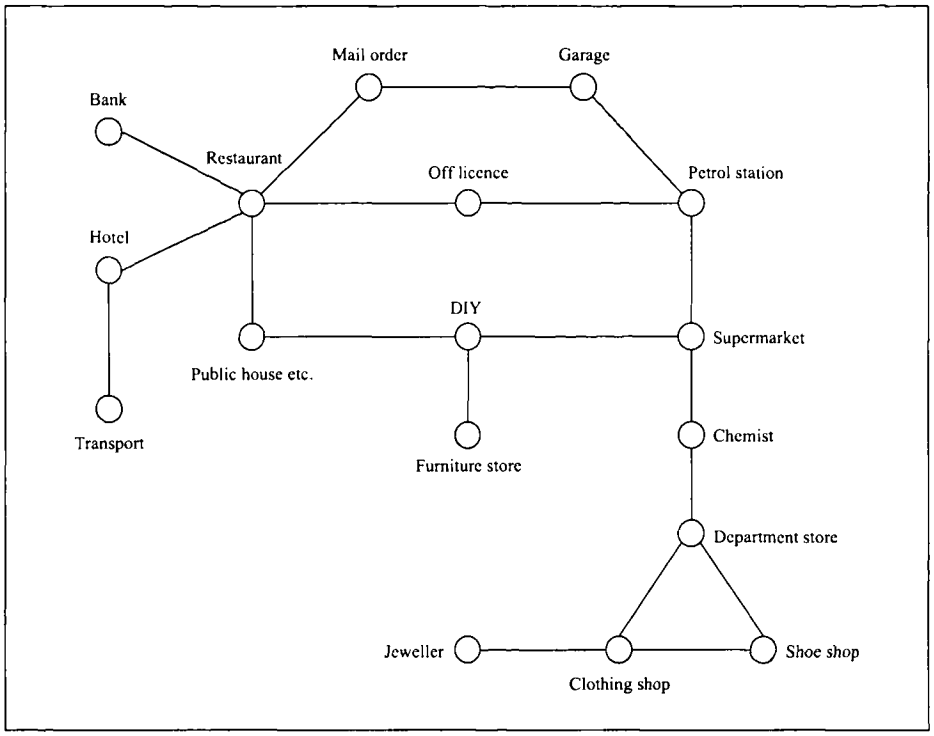
In Figure 8.9 we show a problem we encountered, because most of the values in the scaled inverse correlation matrix are grouped around zero. Little appears to exist in the literature to indicate how, in any particular circumstance, we would be able to determine which edges to leave in, and which to omit. This threshold is critical, because presence or absence of an edge implies either conditional independence, or that variables are related.

Figure 8.9 Values in the scaled inverse correlation matrix



By examination of all elements of the matrix, we settled (subjectively) on a value of 0.095, which gave a good compromise between leaving in those edges which fitted our a priori beliefs and omitting those from a more complete set. Selecting this value led to the model shown in Figure 8.10. The use of automatic selection procedures resulted in graphs that had either hardly any edges, or included most of them. Note that we have not shown all sectors, and have deliberately omitted the five that have an 'other' label, such as 'other shops' and 'other finance'.

Figure 8.10 Significant edge threshold 0.095



The sub-graph connecting the nodes for department stores, shoe shops, and clothing shops is especially interesting here, since each of these nodes refers to purchases related to the same aspect of lifestyle. Hand and Blunt (2001) made the point that *post hoc* explanations can be important in data mining – 'Experience shows that even (perhaps especially) those data mining discoveries which prove to be valuable have a ready explanation – after the fact of their discovery.' Other connections seem less plausible: for example, why should supermarkets and department stores be independent conditionally on chemists? We discuss the last point in Section 8.5.1. We could argue that there are no natural relationships between the sectors, and that we are merely imposing them by a choice of threshold, but this is relevant for monitoring and tracking in a business context, and is immediately usable.

Some of the problems could exist, of course, because of our subjective choice of the threshold value in the conditional independence graph. It would seem to be borne out by other aspects of the data, which we show in Table 8.5, the correlations between spending in the three sectors. This implies that had we selected a slightly different value of threshold, we might have found that department stores and supermarkets were not conditionally independent.

Table 8.5 Correlation coefficients three sectors' spending

	Chemist	Dept store
Dept store	0.31	
Supermarket	0.41	0.28

8.5.1 Spurious links – chemists and supermarkets

One of the edges in several of the graphs we constructed was between chemists and supermarkets. and we believe that this is a relationship that is of little use for marketing purposes, although it had the fourth highest value in the scaled inverse correlation matrix.

Given that a customer spent in a chemist, supermarkets also had a high probability of use, 0.81, although department stores and clothing shops were also high, at 0.82 and 0.81 respectively. We show the full conditional probability matrix in Section 8.6.2. Another feature of spending in supermarkets (although it turns out to be true for any sector as well, which we also discuss in Section 8.6.2) is that, having spent in this sector, customers are more likely to use other sectors disproportionately more.

Thus, the mean number of transactions made in a chemist by a person who uses a supermarket as well is 4.6, compared to 2.0 for those who do not. Expressing this as

a ratio gives 2.25, and this is the highest of all sectors, conditional on spending in the latter. However, there are several reasons why frequent use in chemists is effectively recording a lot of the same information that we can glean by looking at someone's behaviour in supermarkets, as follows.

- Much supermarket spending happens in out of town superstores.
- This limits spending to those with cars (for most practical purposes).
- These people will tend to be slightly more upmarket than the average.
- They will thus be more likely to be credit card holders.
- Many superstore developments have a cluster of small shops on their sites.
- Usually, these smaller shops will be at least a Post Office, a chemist and a newsagent (although many supermarkets sell most of the goods available in the others).
- As the shoppers at these locations are more likely to be credit card holders, they may be more likely to use their cards in chemists as well as supermarkets.

8.6 Sectors used and their probability of use

8.6.1 Introduction

There is another way we could classify sector use – that of the probabilities of use of different sectors, given use in other sectors. This fits naturally into the framework of conditional probabilities. In this section, we examine the conditional probabilities of use of sectors, given use in all others. We extend this to the odds ratio, and compare these two approaches with that of association rules as a way of examining the relationships between sectors. Finally, we show how they are related to the univariate analyses earlier in the chapter.

8.6.2 Conditional probabilities

In Table 8.6 we show the estimated probability of spending in each sector, given that a customer spent in the sector in each row. To make it easier to see what might be happening, we have sorted each sector, both rows and columns, by the column medians increasing from left to right (descriptions of the three letter row and column headings are shown in Table 8.2). The bottom two rows show, respectively, the overall probabilities of use in each sector, and the overall probability of use, but excluding non-spenders.

The spaces at every fifth row are only present as an aid to reading the table. They are not intended to divide the sectors into a grouping of any sort. The grey lines, however, show where there are larger changes in the conditional probabilities, which are more clearly seen in the grey scale plot of Figure 8.11 and where the structure in the table becomes more apparent. The lighter the shading, the higher the value. There are five blocks, divided (in the main) by step changes in most of the columns' conditional probabilities. The column order is very close to that of the groups we described in Figure 7.1, and also Figure 8.8, where we showed the predicted variance (from the NBD) and the skewness of the number of sectors used. In the former, we showed the mean number of transactions per sector and the proportion of customers using that sector.

Table 8.6 Estimated conditional probabilities

	pub	ctn	ofo	off	jew	ban	cin	ole	ofi	fur	tra	mai	trv	hot	gar	che	ele	sho	res	diy	dep	smk	clo	pet	oth	osh
pub		0.19	0.21	0.39	0.37	0.33	0.46	0.48	0.33	0.38	0.55	0.49	0.54	0.68	0.56	0.58	0.58	0.58	0.82	0.73	0.76	0.75	0.79	0.81	0.88	0.94
ctn	0.10		0.20	0.32	0.38	0.32	0.37	0.43	0.38	0.42	0.49	0.52	0.51	0.54	0.59	0.61	0.59	0.56	0.71	0.76	0.74	0.81	0.78	0.81	0.88	0.96
ofo	0.09	0.16		0.35	0.37	0.34	0.35	0.39	0.36	0.42	0.53	0.52	0.47	0.52	0.54	0.67	0.57	0.62	0.73	0.76	0.82	0.88	0.80	0.84	0.85	0.95
off	0.10	0.16	0.21		0.32	0.34	0.33	0.37	0.31	0.36	0.51	0.45	0.43	0.58	0.51	0.59	0.53	0.53	0.76	0.75	0.71	0.81	0.74	0.78	0.85	0.94
jew	0.07	0.15	0.17	0.25		0.31	0.35	0.36	0.35	0.40	0.44	0.46	0.47	0.49	0.53	0.59	0.55	0.60	0.64	0.71	0.78	0.73	0.81	0.72	0.86	0.94
ban	0.04	0.08	0.11	0.18	0.21		0.23	0.23	0.26	0.21	0.33	0.33	0.33	0.34	0.38	0.34	0.37	0.34	0.47	0.50	0.52	0.60	0.57	0.61	0.71	0.78
cin	0.08	0.13	0.15	0.23	0.32	0.31		0.40	0.34	0.36	0.48	0.45	0.53	0.51	0.49	0.49	0.49	0.52	0.68	0.66	0.70	0.68	0.75	0.70	0.85	0.91
ole	0.09	0.15	0.16	0.26	0.33	0.31	0.40		0.35	0.39	0.48	0.48	0.47	0.57	0.52	0.53	0.51	0.54	0.69	0.68	0.73	0.72	0.78	0.76	0.86	0.94
ofi	0.05	0.12	0.13	0.19	0.28	0.31	0.30	0.31		0.32	0.38	0.44	0.46	0.45	0.54	0.45	0.50	0.49	0.55	0.65	0.64	0.67	0.65	0.71	0.84	0.87
fur	0.07	0.15	0.18	0.26	0.37	0.29	0.36	0.38	0.37		0.46	0.46	0.46	0.51	0.54	0.55	0.57	0.58	0.67	0.78	0.78	0.75	0.79	0.75	0.85	0.95
tra	0.07	0.13	0.17	0.26	0.30	0.32	0.35	0.35	0.32	0.34		0.43	0.47	0.58	0.48	0.48	0.49	0.50	0.70	0.64	0.70	0.69	0.70	0.74	0.84	0.91
mai	0.06	0.12	0.15	0.22	0.29	0.30	0.31	0.33	0.34	0.31	0.39		0.40	0.44	0.47	0.47	0.48	0.48	0.55	0.64	0.67	0.66	0.67	0.67	0.85	0.90
trv	0.07	0.12	0.14	0.21	0.30	0.31	0.37	0.33	0.36	0.32	0.45	0.42		0.50	0.46	0.45	0.49	0.49	0.62	0.61	0.70	0.64	0.70	0.66	0.84	0.89
hot	0.08	0.13	0.15	0.27	0.30	0.31	0.34	0.38	0.34	0.34	0.53	0.43	0.48		0.51	0.46	0.51	0.49	0.75	0.66	0.69	0.67	0.72	0.75	0.84	0.91
gar	0.06	0.13	0.14	0.22	0.30	0.32	0.30	0.33	0.38	0.34	0.40	0.43	0.41	0.47		0.45	0.51	0.47	0.60	0.70	0.65	0.69	0.66	0.81	0.84	0.90
che	0.07	0.14	0.18	0.27	0.35	0.30	0.32	0.35	0.33	0.36	0.43	0.46	0.42	0.45	0.47		0.52	0.58	0.63	0.71	0.82	0.81	0.81	0.73	0.83	0.96
ele	0.07	0.13	0.15	0.23	0.31	0.31	0.31	0.32	0.35	0.36	0.41	0.44	0.44	0.47	0.52	0.49		0.51	0.61	0.71	0.69	0.70	0.71	0.72	0.84	0.92
sho	0.06	0.12	0.16	0.23	0.34	0.28	0.32	0.33	0.34	0.36	0.42	0.43	0.44	0.46	0.47	0.55	0.50		0.60	0.69	0.81	0.73	0.81	0.71	0.83	0.95
res	0.07	0.13	0.15	0.27	0.29	0.32	0.34	0.35	0.31	0.34	0.48	0.41	0.44	0.57	0.49	0.49	0.49	0.49		0.66	0.69	0.71	0.71	0.76	0.84	0.91
diy	0.06	0.12	0.14	0.23	0.29	0.29	0.29	0.30	0.32	0.34	0.38	0.41	0.39	0.43	0.49	0.47	0.50	0.49	0.57		0.68	0.71	0.68	0.73	0.81	0.92
dep	0.06	0.11	0.14	0.21	0.29	0.29	0.28	0.30	0.30	0.32	0.39	0.41	0.41	0.43	0.43	0.51	0.45	0.53	0.56	0.64		0.68	0.74	0.66	0.79	0.91
smk	0.05	0.11	0.15	0.23	0.27	0.33	0.28	0.29	0.31	0.31	0.38	0.40	0.38	0.41	0.45	0.51	0.46	0.48	0.58	0.67	0.68		0.70	0.73	0.79	0.91
clo	0.06	0.11	0.14	0.21	0.30	0.31	0.30	0.31	0.30	0.32	0.39	0.40	0.41	0.43	0.43	0.50	0.45	0.52	0.57	0.63	0.73	0.69		0.67	0.79	0.92
pet	0.06	0.11	0.14	0.22	0.26	0.32	0.27	0.30	0.32	0.29	0.39	0.39	0.37	0.44	0.51	0.44	0.45	0.45	0.59	0.66	0.63	0.70	0.65		0.80	0.89
oth	0.05	0.10	0.12	0.19	0.25	0.31	0.27	0.28	0.31	0.27	0.37	0.40	0.38	0.40	0.43	0.41	0.43	0.43	0.53	0.59	0.61	0.62	0.63	0.65		0.86
osh	0.05	0.10	0.12	0.19	0.25	0.30	0.26	0.27	0.28	0.27	0.35	0.38	0.37	0.39	0.41	0.42	0.42	0.44	0.52	0.60	0.64	0.64	0.65	0.65	0.77	
ove	0.04	0.07	0.09	0.15	0.20	0.29	0.21	0.21	0.24	0.21	0.29	0.31	0.31	0.32	0.34	0.33	0.34	0.34	0.42	0.49	0.52	0.52	0.53	0.54	0.67	0.74
exc	0.04	0.08	0.10	0.16	0.21	0.31	0.23	0.23	0.26	0.23	0.31	0.33	0.32	0.34	0.36	0.35	0.36	0.37	0.45	0.52	0.55	0.56	0.56	0.58	0.71	0.79

To illustrate, consider the first row, and the right most column, which shows that, given spend in public houses ('pub'), the probability of using 'other shop' ('osh') is 0.94. The matrix is obviously not symmetric, because we would not expect the probability of (say) using restaurants, given use in chemists to be the same as the probability of use in chemists, given use in restaurants.

One curious feature is immediately apparent – there are hardly any probabilities in the rows that are less than the overall probabilities of using any sector. Only 14 are, in fact, and they are all connected with cash advances ('ban'). This means that (for all practical purposes) given that a customer spent in any sector, she or he is more

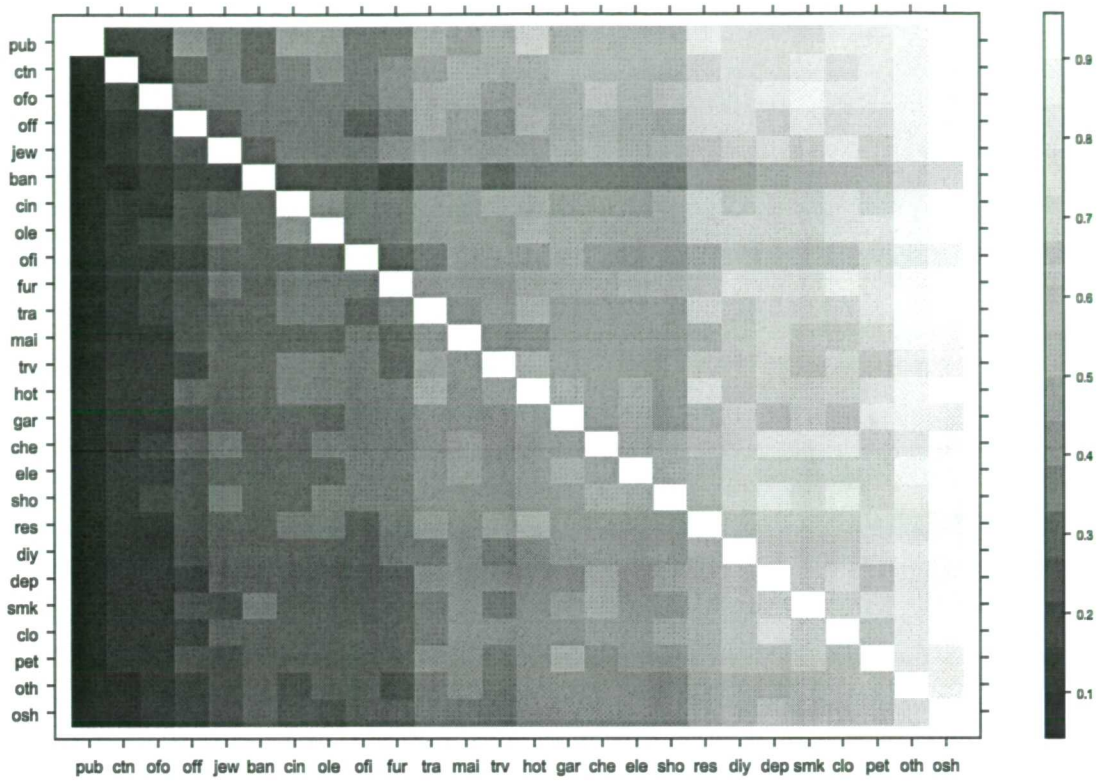
likely to use any of the others. So, we propose that spending in any sector implies higher use of all others, and we speculate that the following structure exists.

		i	
		No use	Used
j	No use	a	b
	Used	c	d

Where $a, d > b, c$, and ***i*** and ***j*** is any pair of sectors.

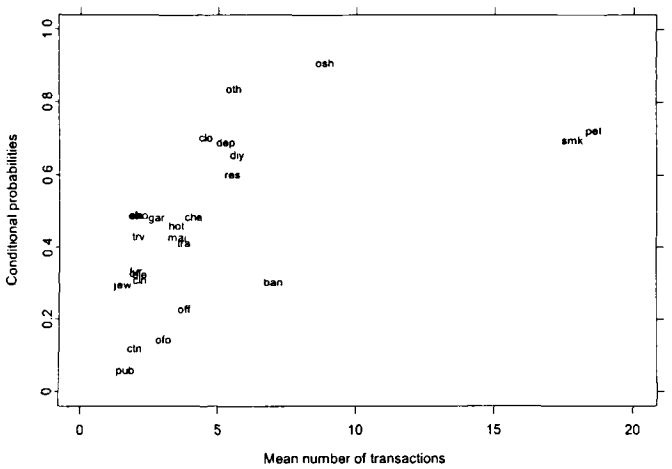
One anomalous row is cash advances ('ban'), for which the conditional probabilities are lower than we might expect from its appearance at this position in the table. In other words, conditional on use for taking cash, all other sectors have a much lower probability of use than would be seen if we conditioned on any other sector.

Figure 8.11 Estimated conditional probabilities



These characteristics could lead us to another type of taxonomy of use, which we show in Figure 8.12. It is the mean number of transactions (for customers using each sector) and the median conditional probability of use by sector (this is the value by which we sorted Table 8.6). Notice the order of sectors is similar to Figure 8.8 The groupings we showed in Table 8.6 (with grey vertical lines) can be seen here too. Cash advances ('ban') are obviously distinct, as are 'other' and 'other shop'. Petrol stations and supermarkets may be anomalous because they have much higher frequencies of use than the other sectors, although their median conditional probabilities of use are very close to the sectors in the same group in Table 8.6.

Figure 8.12 Number of transactions and conditional probabilities



8.6.3 Odds ratio

$P(B \mid A)/P(B \mid \overline{A})$ is known as the odds ratio. If we use the estimated probabilities in Table 8.6, together with their complements, to calculate this statistic, we believe it will prove to be revealing about the relationships between spending in different sectors. A high value will tell us that spending in sector B is much more likely to occur if sector A is used than if it is not. We show the odds ratios in Table 8.7, and, as in Table 8.6, have sorted the rows and columns by the column medians, but this time descending from left to right. Also, the row 'exc' is, as it was in Table 8.6, the overall probabilities of use in each sector (excluding non-spenders). In this case, we have, for each entry in row i and column j, odds ratio = $P(b_j \mid a_i)/P(b_j \mid \overline{a}_i)$, and, as in previous tables, the sectors use the same abbreviations. The vertical lines and the horizontal spaces are shown purely to aid visualisation in Table 8.7, as there are fewer obvious 'break points' comparable to those in the table of estimated conditional probabilities (Table 8.6).

As we have seen on several occasions, cash advances are rather different from most other sectors. Also, the least frequently used sectors tend to have the highest odds ratios, indicated by the fact that they tend to be towards the left of the table.

Another way of expressing this is to say that 'if a customer is happy to use those sectors used by very few people, she or he is prepared to use all of the others as well'. We return to this 'propensity to spend' concept in Section 8.7. Further, we can see some inadequacies in our grouping into trade sectors, particularly in 'other shop'. Eight of the highest values of the odds ratio are in this sector, so we conclude that, ideally, and in future, we would split this sector into more MCC codes.

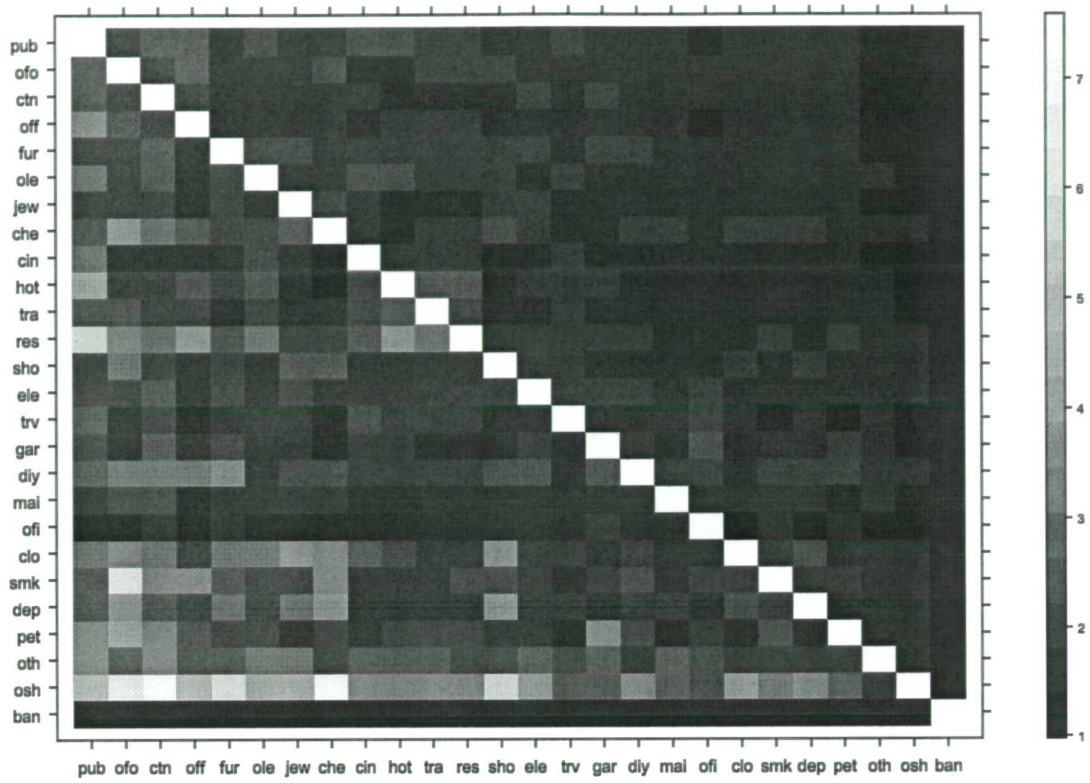
So, a targeting opportunity is to select customers with values of the odds ratios in the middle of the possible range, and to try to encourage them to use some of the more frequently used sectors more regularly.

Table 8.7 Estimated odds ratios

	pub	ofo	ctn	off	fur	ole	jew	che	cin	hot	tra	res	sho	ele	trv	gar	diy	mai	ofi	clo	smk	dep	pet	oth	osh	ban
pub		2.45	2.71	2.78	1.83	2.39	1.93	1.84	2.26	2.23	1.97	2.02	1.72	1.74	1.83	1.66	1.53	1.59	1.39	1.53	1.45	1.48	1.52	1.33	1.28	1.14
ofo	2.67		2.46	2.67	2.15	1.96	2.08	2.28	1.77	1.74	2.00	1.85	1.95	1.80	1.64	1.65	1.65	1.76	1.54	1.60	1.80	1.67	1.63	1.32	1.31	1.20
ctn	2.90	2.40		2.32	2.12	2.19	2.11	2.00	1.84	1.81	1.80	1.78	1.71	1.86	1.75	1.81	1.63	1.76	1.61	1.53	1.61	1.48	1.55	1.34	1.31	1.11
off	3.64	3.03	2.63		1.95	2.00	1.86	2.12	1.72	2.11	2.03	2.08	1.71	1.71	1.52	1.62	1.71	1.56	1.37	1.49	1.71	1.46	1.55	1.34	1.32	1.23
fur	2.24	2.62	2.65	2.11		2.31	2.45	2.05	2.08	1.93	1.87	1.86	2.05	2.04	1.73	1.89	1.91	1.66	1.76	1.74	1.62	1.74	1.53	1.37	1.38	1.00
ole	3.45	2.31	2.77	2.18	2.30		2.10	1.95	2.42	2.25	1.99	1.97	1.84	1.72	1.81	1.78	1.55	1.81	1.63	1.69	1.53	1.56	1.56	1.40	1.35	1.11
jew	2.37	2.42	2.53	1.95	2.38	2.05		2.24	1.97	1.76	1.76	1.71	2.16	1.88	1.79	1.77	1.64	1.65	1.62	1.77	1.53	1.71	1.45	1.38	1.34	1.12
che	2.87	4.14	3.20	2.99	2.49	2.32	2.94		1.99	1.79	1.90	1.96	2.54	2.01	1.70	1.70	1.86	1.85	1.66	2.05	2.10	2.18	1.60	1.42	1.49	1.04
cin	3.13	2.01	2.16	1.82	2.07	2.42	2.01	1.74		1.92	2.01	1.92	1.76	1.64	2.16	1.61	1.49	1.63	1.59	1.60	1.41	1.47	1.40	1.38	1.31	1.09
hot	4.53	2.31	2.55	2.91	2.26	2.78	2.04	1.77	2.25		2.94	2.78	1.79	1.93	2.16	1.91	1.60	1.66	1.72	1.64	1.49	1.58	1.70	1.44	1.37	1.11
tra	2.98	2.77	2.39	2.53	2.06	2.23	1.96	1.82	2.26	2.72		2.22	1.77	1.73	1.99	1.65	1.52	1.59	1.53	1.53	1.51	1.57	1.58	1.41	1.35	1.19
res	6.33	3.65	3.37	4.27	2.70	3.07	2.38	2.35	2.91	4.13	3.12		2.07	2.11	2.18	2.03	1.83	1.64	1.63	1.81	1.84	1.75	1.96	1.55	1.46	1.19
sho	2.58	3.08	2.40	2.17	2.60	2.21	2.92	2.65	2.07	1.85	1.88	1.83		1.96	1.85	1.70	1.81	1.73	1.80	2.10	1.77	2.19	1.56	1.43	1.48	0.99
ele	2.62	2.59	2.84	2.14	2.55	2.00	2.32	2.06	1.87	1.99	1.83	1.85	1.95		1.88	2.03	1.91	1.79	1.90	1.61	1.64	1.59	1.61	1.46	1.41	1.12
trv	2.69	2.05	2.33	1.72	1.91	2.04	2.04	1.67	2.57	2.11	2.03	1.81	1.77	1.80		1.57	1.42	1.54	1.90	1.56	1.36	1.58	1.35	1.41	1.31	1.10
gar	2.39	2.21	2.72	1.97	2.28	2.11	2.14	1.73	1.83	1.98	1.74	1.80	1.70	2.04	1.63		1.87	1.71	2.23	1.45	1.59	1.42	1.98	1.46	1.35	1.18
diy	2.86	3.32	3.32	3.19	3.72	2.21	2.60	2.56	2.03	2.01	1.89	2.04	2.34	2.59	1.67	2.47		1.84	1.97	1.77	2.10	1.83	2.02	1.52	1.58	1.05
mai	2.10	2.33	2.40	1.79	1.83	2.05	1.85	1.83	1.80	1.66	1.63	1.49	1.68	1.74	1.55	1.66	1.52		1.74	1.44	1.43	1.50	1.39	1.47	1.32	1.08
ofi	1.56	1.73	1.87	1.43	1.82	1.67	1.68	1.55	1.63	1.61	1.48	1.42	1.63	1.70	1.77	1.92	1.50	1.63		1.33	1.40	1.32	1.45	1.37	1.24	1.11
clo	3.37	3.60	3.10	2.48	3.43	3.13	3.89	3.75	2.67	2.29	2.11	2.20	3.68	2.13	2.12	1.76	1.90	1.81	1.65		2.10	2.56	1.67	1.50	1.67	1.18
smk	2.66	6.50	3.82	3.87	2.69	2.33	2.40	3.87	1.91	1.88	2.00	2.23	2.51	2.17	1.63	2.03	2.28	1.76	1.84	2.07		1.95	2.17	1.48	1.61	1.37
dep	2.89	4.15	2.67	2.26	3.31	2.45	3.30	4.16	2.11	2.10	2.16	2.07	3.90	2.05	2.12	1.69	1.93	1.90	1.62	2.49	1.94		1.56	1.46	1.62	0.99
pet	3.59	4.34	3.60	2.94	2.46	2.63	2.20	2.25	1.97	2.58	2.34	2.61	2.07	2.21	1.64	3.47	2.30	1.74	2.05	1.70	2.28	1.60		1.56	1.55	1.30
oth	3.50	2.92	3.48	2.88	2.78	3.07	3.05	2.47	2.88	2.67	2.68	2.62	2.49	2.70	2.58	2.69	2.13	2.91	2.62	1.90	1.88	1.84	1.97		1.65	1.24
osh	5.48	6.87	7.45	5.34	6.76	5.03	5.10	7.59	3.62	3.60	3.52	3.48	6.06	4.07	2.79	3.02	3.76	2.92	2.37	3.89	3.38	3.55	2.75	2.05		1.23
ban	1.20	1.27	1.15	1.29	1.00	1.13	1.14	1.04	1.10	1.10	1.19	1.15	0.99	1.11	1.10	1.16	1.03	1.08	1.11	1.12	1.22	0.99	1.17	1.10	1.07	
exc	0.04	0.10	0.08	0.16	0.23	0.23	0.21	0.35	0.23	0.34	0.31	0.45	0.37	0.36	0.32	0.36	0.52	0.33	0.26	0.56	0.56	0.55	0.58	0.71	0.79	0.31

Figure 8.13 shows a grey scale plot of the odds ratios in Table 8.7. Again, cash advances (the bottom row and right most column) are clearly different from most other sectors. There is a much smoother gradient (using the term in an informal, rather than strictly mathematical sense) from bottom left to top right, than we saw in the plot of conditional probabilities. Use of the card for cash advances ('ban'), alone of all the sectors, hardly depends on the use of other sectors. This is in contrast to all of the others, where we see the phenomenon we have already discussed – use of most sectors implies greater probability of use in any particular sector.

Figure 8.13 Estimated odds ratios



8.6.4 Association rules

In much association rules literature the emphasis is more concerned with *pattern recognition* than *model building*, a distinction we have made several times. The algorithms produce 'sets of patterns that are local in the sense that they apply to specific regions of the p-dimensional space' (Smyth, 2000). However, our spending data is similar to a typical application of association rules – *market basket analysis* – where the term comes from supermarket shopping baskets, and the purchasing patterns seen there.

We will demonstrate that, although algorithms have been developed to produce association rules, and allow the rapid selection of those with high *confidence* and *support*, the insight they give for a data set such as ours is less than can be gained

using traditional statistical methods. This phenomenon was noted by Berry and Linoff (2000) 'we don't often get much mileage out of association rules'; and later, 'one of the difficulties with association rules is that there are too many of them'. Smyth (2000) notes that 'it is difficult to find any specific published reference which describes a successful application to a real problem'. He goes on to say that 'the method is primarily a computationally efficient data analysis technique for massive transactional data sets'. Indeed, for our data set, the association rules that we describe ran in minutes on a personal computer, compared to the hours taken by *k*-nearest neighbour methods (described in Chapter 5).

As we described in Chapter 1, Padmanabhan and Tuzhilin (1999) gave details of a study which resulted in 20,000 rules, and that the 'the interpretation and evaluation of the discovered rules could be a highly resource-consuming exercise'. These authors, and others, considered 'interestingness' measures as a way of assessing association rules: for example Dong and Li (1997), Freitas (1998, 1999), Sahar (1999) although a literature search finds many more. We believe that the aim of finding an automatic way to assess interestingness will always be fraught with problems, and that domain knowledge will always be much more important. We have already mentioned cash withdrawals, as one example of this, and a further illustration will be given in Section 8.7.

Association rules were defined in Hand, Blunt and Bolton (2001) as follows.

'Let V be a set of attributes, corresponding to the set of all possible items which can exist in a database. A *transaction*, T, is a subset of V. For convenience, we will refer to 'the set of transactions in the database', although technically, since the database contains items, we should refer to the set of transactions in the set of subsets of V.

'An *association rule*, R, is a pair (A, B), where A is a subset of V, and B is (often) a single element of V, not in A. A is often called the *antecedent* of the rule, and B the *consequent*. A *transaction*, T, is then said to *satisfy* a rule $R = (A, B)$ if the elements in R are all in T. For any C a subset of V, P(C) represents the proportion of transactions which include C as a subset. That is, $P(C) = \mathit{frequency}(T \mid C \subseteq T)$.'

Association rules are often denoted by $A \Rightarrow B$ (see, for example, Agrawal et al (1993a), Dong and Li (1997), Pasquier et al. (1999), Han and Kamber (2001) and many others), but this seems to imply causality, which may not be justified. In general, association rules do not allow us to deduce this sort of relationship, but simply that the 'rules' measure the occurrence of certain conjunctions of events being seen in the data. Following on from Hand, Blunt and Bolton (2001), we propose the following definitions:

1. Support: $S = P(A)$
2. Confidence: $C = P(B \mid A)$
3. Lift: $L = P(B \mid A)/P(B)$

The odds ratio is as described in 8.6.3, and in the tables that follow, we have abbreviated it to OR.

Berry and Linoff (2000) describe lift as 'the likelihood of finding the right-hand product in a basket known to contain the left-hand product, to the likelihood of finding the right-hand product in any random basket', having earlier defined an association rule as 'left-hand side implies right-hand side'. In this case, the two 'sides' of the rule are respectively the antecedent and consequent of the rule, so their definition of lift is the same as ours.

Let **X** be a 3,972 × 26 element matrix where, for each customer i in sector j, we have the following

$$x_{i,j} = \begin{cases} 1 \text{ if spend in sector } j \geq 0 \\ 0 \text{ if spend in sector } j = 0 \end{cases}$$

Mannila (1997) says 'In large retailing applications the number of rows might be 10^6 or 10^8, and the number of columns around 5,000. The frequency threshold σ typically is around 10^{-2} - 10^{-4}. The confidence threshold θ can be anything from 0 to 1.' In our case, our data set is relatively small (by data mining standards, and we discussed this in Chapter 3), but if we wanted to model the entire Barclaycard file, it would have ~10^7 rows, and several hundred columns if we included all data at their most detailed level. In our sample, we could easily extend the number of columns, because we could use the individual Merchant Category Codes, rather than our sector level grouping. We drew the distinction between *patterns* and *models* in Chapter 1, and if it was our aim to search for patterns in the data, we might need to use the whole data set at the most granular level of detail. Our objective here is to deduce whether or not association rules can give any insight into suitable *models*, rather than to find all small *patterns* in the data.

Bradley et al. (1999) briefly discuss association rules and point out several key features – the number of rules grows exponentially; that many are going to be too sparse to satisfy given support and confidence threshold; and that rules shouldn't be viewed as statements about causal effects in the data.

Since the original Agrawal et al. (1993a) work was published, there have been many others (for example Agrawal et al. 1993b, Agrawal and Srikant, 1994, Boulicat et al.,

1998, Han and Fu, 1999, Pasquier et al., 1999, Hipp et al., 2000, Witten and Frank, 2000, Han and Kamber, 2001). We do not seek to provide a comprehensive review of this method, because we see it as another data mining tool, rather than one to be used to the exclusion of statistical methods.

Han and Kamber (2001) define a rule as 'strong' if it satisfies minimum support and confidence thresholds, then note that 'strong rules are not necessarily interesting', but we believe that some high levels of confidence can be misleading. Table 8.8 shows the first five association rules from our spending data, ranked by confidence and support; if we were to show the next few thousand we would see that the first 2,303 have 'other shop' as a consequent. Partly this is because of the phenomenon we observed in Section 8.4.5, which was caused by the grouping of several MCCs into one trade sector. We have shown the other statistics necessary to assess the effectiveness of the rules. We have used the Apriori algorithm, developed by Agrawal and others (Agrawal et al., 1993a, Agrawal et al., 1993b, Agrawal and Srikant, 1994).

Table 8.8 First five association rules, ranked by confidence and support

Antecedent	Supp.		Conseq.	Conf.	P(B)	$P(B \mid \overline{A})$	OR	Lift
Dept, petrol, rest, shoe, smkt	0.137	$\Rightarrow$	Other shop	1.000	0.791	0.758	1.320	1.265
Dept, petrol, rest, shoe, smkt, other	0.127	$\Rightarrow$	Other shop	1.000	0.791	0.760	1.315	1.265
DIY, elec. goods, shoe, smkt,	0.125	$\Rightarrow$	Other shop	1.000	0.791	0.761	1.314	1.265
Dept, DIY, rest, shoe, smkt, other	0.124	$\Rightarrow$	Other shop	1.000	0.791	0.761	1.314	1.265
Clothing, dept, DIY, rest, shoe, other	0.124	$\Rightarrow$	Other shop	1.000	0.791	0.761	1.314	1.265

In Table 8.9, we show the first five rules that result from omitting the 'other' sectors, and again sorted by confidence and support. Arbitrarily taking the first 1,000 rules, we see that the eight sectors in the table accounted for 82% of appearances in either

the antecedents or the consequents of these rules. Taking the rule with the highest lift value in this subset (2.34), we see the following.

Use in {chemist ∩ DIY ∩ restaurant ∩ shoe shop} ⇒ use in {dept. store ∩ supermarket}

The only sectors missing from the most commonly used ones are petrol stations and shoe shops.

Table 8.9 Excluding the 'other ...' sectors

Antecedent	Supp.		Conseq.	Conf.	P(B)	$P(B \mid \overline{A})$	OR	Lift
Chemist, clothing, dept, DIY, petrol, rest	0.114	⇒	Smkt	0.960	0.556	0.504	1.906	1.728
Chemist, clothing, DIY, rest, shoe	0.107	⇒	Smkt	0.952	0.556	0.508	1.874	1.714
Chemist, clothing, DIY, petrol, rest	0.128	⇒	Smkt	0.952	0.556	0.497	1.914	1.713
Chemist, dept, DIY, rest, shoe	0.108	⇒	Smkt	0.950	0.556	0.508	1.872	1.711
Chemist, dept, DIY, petrol, rest	0.128	⇒	Smkt	0.950	0.556	0.498	1.908	1.710

If we examine all of the 7,471 rules produced (at confidence threshold set to 0.6, and support to 0.1) six sectors comprise 60% of all appearances in the antecedents, and 89% of all appearances in the consequents. They are clothing shops, department stores, DIY, petrol stations, restaurants and supermarkets.

8.6.5 Visual assessment of association rules

Hofmann and Wilhelm (2001) describe graphical methods for the selection of interesting association rules. We had tried graphical methods on our data, and saw a similar characteristic, where points appeared as 'rays exiting from the origin'. However, we found that this simply transferred the selection of appropriate support and confidence thresholds to be one of selecting the optimum region on a two dimensional plot. We have as little guidance for the latter as the former.

Their 'difference of confidence', *doc* for an association rule $A \Rightarrow B$ is as follows.

$$doc(A \Rightarrow B) = \text{Conf.}(A \Rightarrow B) - \text{Conf.}(\neg A \Rightarrow \neg B)$$

Where Conf. is the confidence of the rule.

Hofmann and Wilhelm (2001) describe several novel graphical methods for assessing the 'quality of an association rule'. We have not pursued these ideas further, because, as the authors say, these methods assume 'that the vast number of rules generated by the standard association rule algorithms has already been reduced to a manageable size by some automatic filtering tool'. We pointed out some problems that can arise with automatic detection in Chapter 7, and that could be an issue here too. Their methods would seem to help when the number of rules has been reduced, but that still leaves open the question of how to select a subset of interesting rules.

8.6.6 Conclusions about association rules

We saw that a few sectors appear frequently in the first several hundred rules. As Berry and Linoff (2000) found in their work, there seems to be little mileage in pursuing this approach. All the association rules permit, at least for our data, is the extraction of combinations of sectors that are often used together. In other words, the technique produces many *patterns* in the data, but little insight into relationships, or *models* of the whole data set. We drew this distinction in Chapter 1. The conditional probability and odds ratio tables gave, much more succinctly, the strengths of the relationships between sectors. Smyth (2000) suggests that association rules can be used for modelling if the rules are viewed as 'constraints on

a large p-dimensional contingency table' and then fitted by 'joint probability models which are consistent with these constraints'.

Examination of different combinations of odds ratios and estimated conditional probabilities allowed us better to deduce which pairs of sectors are used most in combination. Also, the structure in Table 8.6 and Table 8.7 guides us towards groups of customers who use their cards in particular sectors, and which other sectors we may be able to persuade them to use, if we can provide suitable incentives. So, for example, let the probability of spending in *restaurant* = $P(r)$, and in *off licence* = $P(o)$, then we know, from Table 8.7

$$\frac{P(o \mid r)}{P(o \mid \bar{r})} = 4.27$$

and from Table 8.6

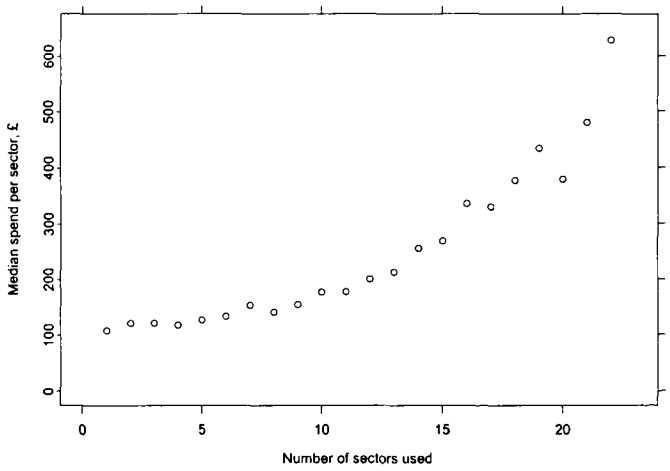
$$P(o \mid r) = 0.27 \text{ and } P(o) = 0.16$$

Such an incentive can now be targeted at customers who use their cards in restaurants, but not in off licences, and it would be to try to encourage spending in the latter. This ought to work better than an offer sent to a random group of customers because the two types of purchase could be linked by similar elements of lifestyle. We return to this issue in the next section. Another incentive could be offered to customers who use both, to spend on higher priced items. Each uplift in turnover does not have to be large for each individual, but added across several tens of thousands of customers, would have a massive impact on the business.

8.7 Number of sectors and amounts spent

Figure 8.14 shows the relationship between the median spend per sector and the number of sectors in which the card is used. The shape of this curve was unexpected: one might have expected that use of more sectors would lead to smaller spend per sector, on the grounds that each individual's pot of money is finite. That is, one might have expected the slope of the curve to decrease to the right. Being wary of causal interpretations, it appears from this diagram that those who use the card in more sectors spend disproportionately more.

Figure 8.14 Median spend per sector against number of sectors used



In Figure 8.15 we show the relationships between the number of sectors used and the median amount spent in each one (we have transposed the axes from the orientation used in Figure 8.14 for easier visualisation). Each panel shows one sector, and has, on the vertical axis, the number of sectors used, grouped as follows: less than or equal to 3 ('le3'), 4, 5, ..., 20, 21, greater than or equal to 22 ('22+'). On the horizontal axis, we show the median total amount spent, and each panel is drawn to

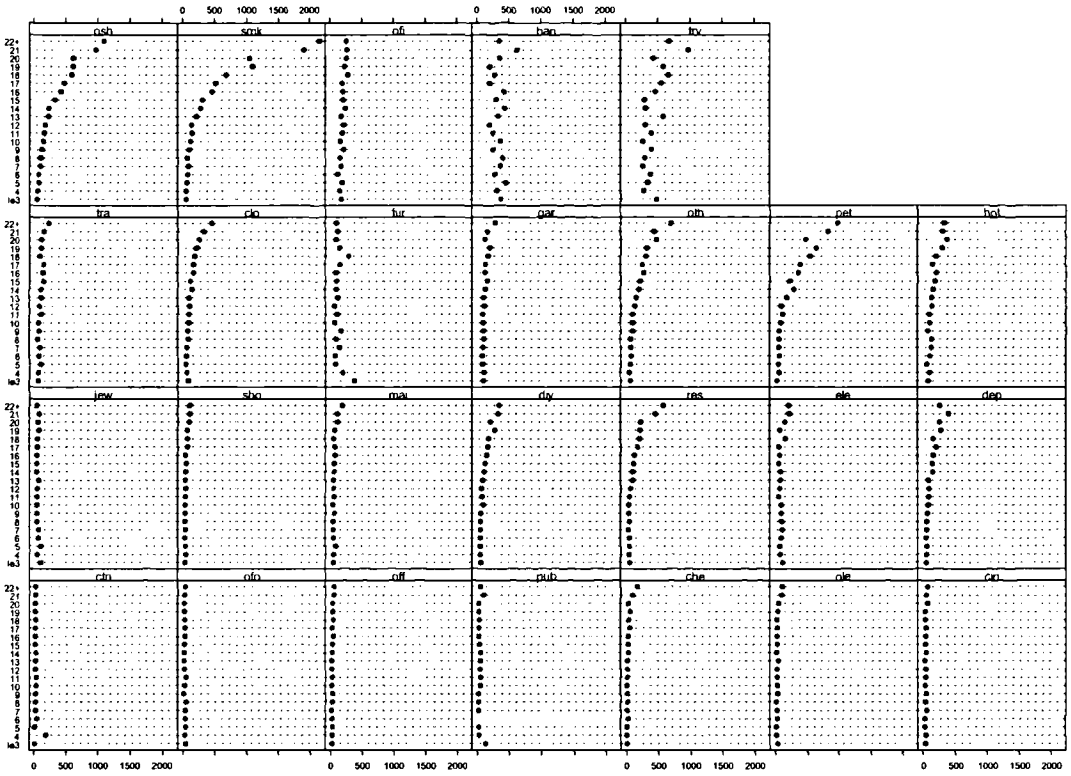
the same scale. We have grouped the highest and lowest number of sectors used, because the cells at the extremes often had small numbers of customers.

At this stage we have used this simple summary measure to avoid showing too much data, which could easily be the case if we included the amounts spent in each sector, for each number of sectors used (see Chapter 7 for illustrations of how difficult it can be to interpret dense scatter plots). In section 8.7.1, we will look at some of these distributions in more detail, because we strive to lose as little information as possible, and will investigate them to see if we can deduce any patterns that the summary measures might conceal. The abbreviations we have used to label each panel are as shown in Table 8.2.

We would expect random variation, as amounts approach zero, to be to the right on each panel, because each distribution is positively skewed. By and large, spending on a credit card cannot be negative. We say 'by and large' because a customer may be given a refund by means of a credit voucher, which appears as a negative transaction, but we have removed these from our sample. These typically amount to about 1% of transactions.

It is apparent that some sectors see much higher levels of activity, and particularly supermarkets. Many sectors, especially those where the transaction value is small and frequent, have a similar pattern, which is a slope from the bottom left to top right. Expressing that in a different way: for most sectors, the more sectors that people use, the more that they spend in each sector. Notable exceptions are furniture stores, jewellers, cash advances; and some show little variation until the number of sectors used is around 17 or more – these are travel agents, cash advances and 'other leisure'.

Figure 8.15 Median spend per sector (horizontal axis of each panel) by the number of sectors used (vertical axis of each panel) and each sector

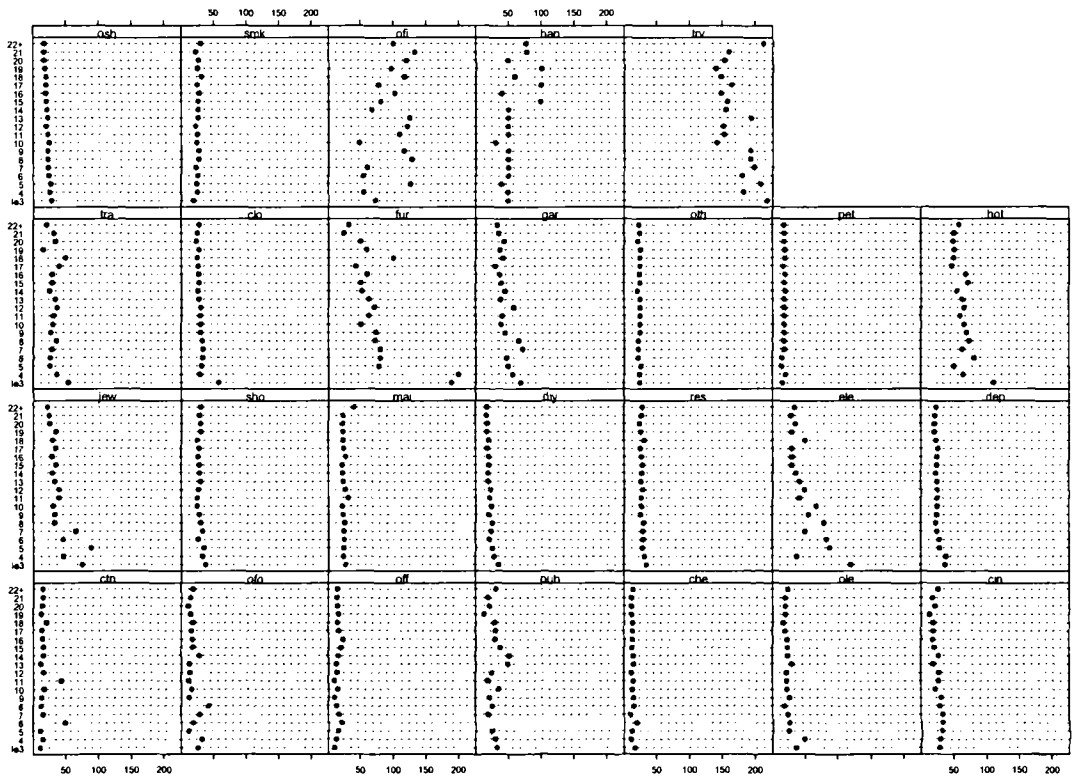


Analysis of many variables simultaneously can sometimes reveal structures which are unexpected and concealed by simple univariate analyses, but when more than two variables are involved, the possibilities for concealed (and possibly confusing) effects are greater. A classic example is Simpson's paradox, which we discussed in Chapter 1. Hand and Blunt (2001) noted that 'contingency tables can be used to explore the structure of relationships between categorical variables, but they become difficult to use with four or more variables'.

We showed a graphical model in Section 8.5, and this method can be useful when there are many variables, but graphical models may be difficult to interpret if the

graph has many edges. The structure of the graph we showed in Figure 8.10 is simple, but that was only because we reduced the number of potential edges quite markedly. We did this by selecting an appropriate threshold (for the selection of significant edges) by eye and prior expectations of likely behaviour. In this instance, domain knowledge was essential. There is no ideal way of reducing a complicated multi-dimensional structure to one that is simple and low dimensional, but which reveals the underlying structure (Bradley et al., 1999, Hand and Blunt, 2001).

Figure 8.16 Median transaction value in each sector by panel



Median transaction value in each sector by panel

In Figure 8.16 we see that the median transaction value in supermarkets is essentially independent of the number of sectors in which people use their card. This is the case for many sectors, although there are a few exceptions. The latter are usually those

that tend to have high value (and often infrequent) transactions, such as furniture stores ('fur') and jewellers ('jew'). There is an important consequence of this for the business – if it seeks to generate extra spending (which most do) then the most important thing is to encourage customers to make more transactions. These data suggest that an alternative approach, that of seeking to raise the value of each individual transaction (i.e. trying to incentivise people to spend more each time they use their cards), would not work as well.

One feature of this plot might appear unusual, but it is an artefact of the data, and it is for cash advances, (as before, we have abbreviated this sector to 'ban'). Most cash is taken from automated teller machines (ATMs), and the values are thus highly constrained, usually at multiples of £10, and no others. This would not necessarily be the case for cash taken abroad (around 8% of the total number, 11% of the total amount in our sample), where the value of the local currency would be converted to sterling when the transaction is posted to the account. The result of this is that 87% of cash withdrawals are for whole pound amounts, and further, most of these are for multiples of £10.

8.7.1 Distribution of transactions by number of sectors used

One of the advantages about using a simple summary measure such as the median is that it allows us to show matrix plots like the previous two, and be able to discern structure about sectors and how they are used. The main drawback though, as should be obvious from earlier in this chapter, is that most of our distributions are markedly skewed.

First, consider petrol stations and supermarkets, both of which have many transactions, so plots of their transactions are likely to be well populated at all possible numbers of sectors. In Figure 8.17, we show the distribution of annual spending in petrol stations, compared to the 'pet' panels in Figure 8.15 where we showed only the median value. The bar plot shows the number of customers who used each number of sectors. Figure 8.18 shows similar data but for supermarkets. These plots form a link between those that we showed in Figure 8.7, the number of sectors used by sector, and the matrices of plots in Figure 8.15 and Figure 8.16. The increasing curve of the median conceals the fact that the majority of users still have a low spend for each number of sectors in which the card is used (notice also the different scales).

Figure 8.17 Petrol station spend, with the number of customers at each number of sectors

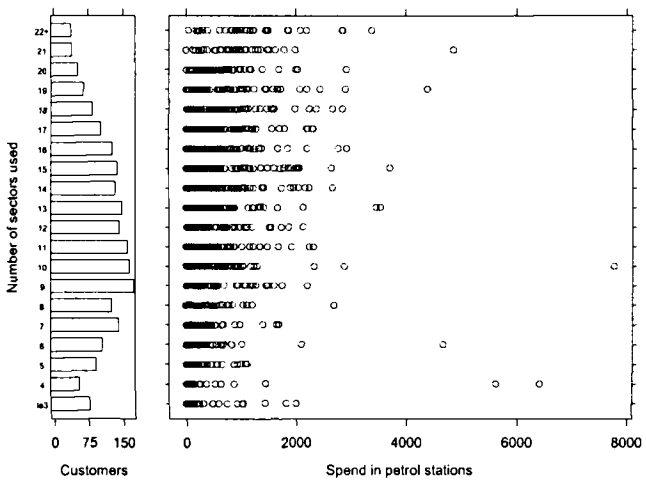
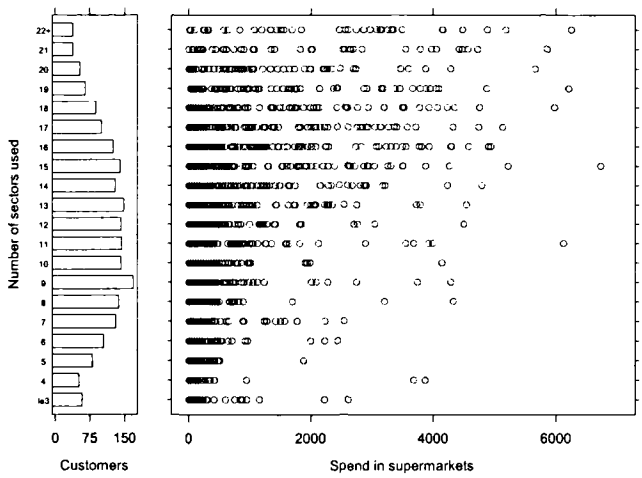


Figure 8.18 Supermarket spend, with the number of customers at each number of sectors



Similar patterns are seen for most sectors.

8.8 Principal components, and cluster analyses

Principal component analysis seeks linear combinations of a data matrix $\mathbf{X}$ with maximal (or minimal) variance, and cluster analysis seeks to group the data into homogeneous clusters. Given the results of the data that we have described throughout this chapter, we did not expect much to be revealed by either PCA or cluster analysis, because the data do not appear to be related in any simple linear way, nor able to be easily partitioned.

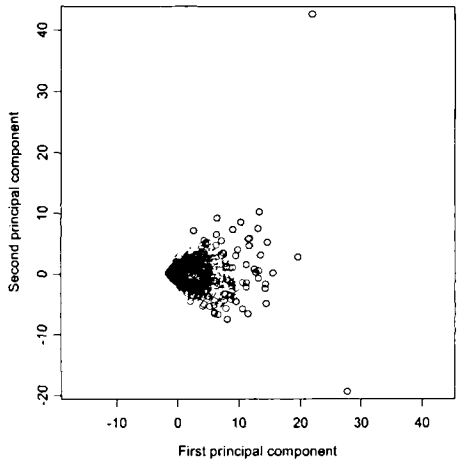
The reliance of PCA on variance may lead to problems (Ripley, 1996) because the variances of our data by sector are very different. Let $\mathbf{X}$ be a $3{,}972 \times 26$ element matrix where $\mathbf{x}_i$ is the amount spent in sector i by each customer. In this case we have $\text{var}(\mathbf{x}_{\min}) = 584.6$ (in public houses) and $\text{var}(\mathbf{x}_{\max}) = 1{,}499{,}788$ (for mail order). Alternatively, if $\mathbf{X}$ is a $3{,}972 \times 26$ element matrix where each $\mathbf{x}_i$ is the total

number of transactions in sector i by each customer. In this case we have $\text{var}(\mathbf{x}_{\min}) = 0.166$ (in public houses) and $\text{var}(\mathbf{x}_{\max}) = 469.6$ (supermarkets).

To explore this method further, we show, in Figure 8.19, the first two principal components for the first definition of $\mathbf{X}$, where each $\mathbf{x}_i$ is the total amount spent, and we have scaled the variables to have unit variance. This, of course, imposes a weighting on the variables, and we do not know if this is a reasonable course of action. If we had not done so, the plot would have been dominated even more by high spending customers. However, it illustrates the problems that can arise with such 'messy' data.

There are two points which are obviously distant from the rest, but removing these does not reveal much more structure. If we remove these, then we find that several more points might appear to be outliers. We could omit those as well, but this 'shaving' approach would gradually lead to the removal of around a third of the data, by which stage the exercise would become pointless if our aim is to reduce dimensionality in the data.

Figure 8.19 PCA for amounts spent



We will not show the results of the cluster analyses that we performed, because the results revealed as little as PCA. Having seen the results of the latter, this was what we expected: there were no discrete clusters, regardless of how many clusters we used to try to partition the data.

8.9 Conclusions

8.9.1 Characteristics of spending behaviour

The curve and analysis in Section 8.7 suggest that a useful measure of behaviour is captured by the notion of 'propensity to use the card'. This property is characterised by the increasing amounts spent per sector, as the number of sectors used by the customer rises. However, finding a suitable measure for this is likely to prove difficult, and some simple relationships between pairs of sectors were found. We have discussed, several times throughout this work, the importance (to a bank) of having rules that can be interpreted easily, and are hence almost certain to be simply expressed.

Contingency tables can be used to explore the structure of relationships between categorical variables, but they become difficult to use with four or more variables. Graphical models can be useful when more variables are involved, although ease of interpretation depends on the graph not having too many edges. We found that for our data, the number of edges was not at all easy to determine. There is no simple technique for reducing a high dimensional structure to a lower one, and which retains the interesting characteristics we seek. Association rules, while being a computationally efficient way of discovering local patterns, are a verbose way of showing the most commonly used combinations of sectors, and there is no easy way to combine many rules into a cohesive structure.

We might seek, in future work, to split our trade sectors into more levels of detail, although all the problems we have just mentioned might become even more apparent. There would be many more dimensions, and one problem we have not discussed at all – many sparse data elements, as the number of potential variables rises.

8.9.2 Conclusions about a taxonomy of spending behaviour

We have considered four ways of modelling customers' use of different sectors, as follows.

1. number of transactions per sector per customer
2. transaction value in each sector
3. number of sectors used conditional on use in each sector
4. conditional probabilities and odds ratios

The first, third and fourth of these are interesting from the business' perspective, because increasing customers' use of their cards is a desirable objective for the business. The second, although exhibiting interesting patterns, is of less use, because

much of the structure occurs as a result of pricing policies in different merchants, rather than as a direct result of customers' behaviour. We are less likely to be able to change this than the behaviour of our customers.

To develop a taxonomy of sector use we need to be able to allocate each sector to a group on the basis of some classification schema that we can devise. We do not have, from these analyses, an obvious partition of sectors into different classes, because many of the discriminating variables are continuous. There are some sectors, particularly supermarkets, petrol stations and cash advances that are distinct, and could be allocated to 'classes', but that would leave all of the other sectors in a third class, which seems unsatisfactory.

We can, however, use different measures to allow us to describe sectors and the differences between them. This would be true wherever we partition any of the variables we have described in Sections 8.2, 8.3 and 8.4. So, we propose the following – there are two potential variations of taxonomies.

The first is by transactions in each sector, and would group by rounding. Sectors where transactions are rounded by choice (e.g. petrol stations), sectors where transactions are rounded by pricing policies (e.g. department stores), sectors where there is no rounding (e.g. supermarkets) and sectors where there are a small number of anomalous spikes (e.g. travel agents). Each of these fits into the mixture distribution framework we described in Section 8.4.

The second is to partition along one of three continua – using parameters from the NBD, conditional probabilities or odds ratios. None of these continua will yield a discrete set of categories, but use of any of these variables allows us to measure

position along a continuum of various types of use. We would measure the success of any incentives – or other marketing activity – against the appropriate measures on each continuum. We return to this notion of *segmentation* for convenience in Chapter 10. So, for example, if we were seeking to encourage particular customers to use more sectors, we would select use in a small number of sectors, based on the odds ratios and conditional probabilities, and then measure the shifts against the number of sectors used by our target groups.

Chapter 9

9 Transaction data – predictive models

9.1 Introduction

We described, in Chapters 4 and 5, the prediction of repayment behaviour. In this chapter, we seek to achieve similar results, but this time applied to spending. As mentioned in Chapter 7, spending in different sectors is seasonal. For this exercise, ideally we need to look at peoples' spending in one 12 month period, and predict spending in the following year, rather than restricting ourselves to the 19 months we used in the earlier chapters. To enable us to do this, we took a supplementary extract of data of customers' transactions, as described in Chapter 3, which was in addition to the original January 1996 – July 1997 period of account history data.

The result of this, as described in Chapter 3, is that a small number of customers were 'lost' from our sample – 145 from the design set, and 146 from the test set. All of the analyses in this chapter are performed on the slightly smaller samples, respectively 3,827 and 3,826 customers in each, rather than the 3,972 in each on which the analyses in Chapters 4 - 6 were undertaken

9.2 Predicting peoples' total spend

9.2.1 Using the sectors in which they spend

We described, in Chapters 7 and 8, how peoples' spending is related to the number of sectors they use, and we speculate that it might be possible to predict their total spending from the number of sectors they use. We saw that the relationship was non-linear, and that use of more sectors was accompanied by disproportionately

higher levels of spending. Let $\mathbf{X}$ be a 3,972 × 26 element matrix where, for each customer i in sector j, we have the following

$$x_{i,j} = \begin{cases} 1 \text{ if spend in sector } j \geq 0 \\ 0 \text{ if spend in sector } j = 0 \end{cases}$$

Then, form the model $\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \varepsilon$, where $\mathbf{y}$ is the total amount spent by each customer in the year. Some summary results from this are as shown in Table 9.1, and the adjusted $R^2 = 0.56$.

Table 9.1 Summary results from linear regression

	Value	Std. Error	t value	Pr(>|t|)
Intercept	-506.7	78.0	-6.50	0
Other food	1057.8	129.4	8.18	0
CTN	937.5	138.5	6.77	0
Furniture	863.2	94.1	9.17	0
Hotel	859.8	89.1	9.65	0
Public houses	762.1	185.5	4.11	0
Other finance	751.6	85.4	8.80	0
Travel agent	730.5	82.6	8.84	0
Off licence	715.0	105.5	6.78	0
Other leisure	695.3	93.6	7.43	0
Electrical	638.7	82.6	7.73	0
Garage	622.1	83.1	7.49	0
Mail order	614.9	80.2	7.67	0
Transport	570.6	86.4	6.61	0
Jewellery	551.4	95.4	5.78	0
Shoe shop	410.4	86.4	4.75	0
DIY	410.3	83.4	4.92	0
Cinema	354.0	92.0	3.85	0
Restaurant	325.2	87.9	3.70	0
Bank	299.5	77.1	3.88	0
Chemist	266.1	89.8	2.96	0
Petrol	180.9	84.4	2.14	0.03
Department store	168.5	85.9	1.96	0.05
Other	144.6	85.1	1.70	0.09
Clothing	141.4	86.4	1.64	0.10
Supermarket	84.0	84.8	0.99	0.32
Other shop	-497.0	98.8	-5.03	0

We have taken the slightly unusual step of ordering the coefficients (with the exception of the intercept), to demonstrate more easily the similarity between this model and the table of the odds ratios. The order is very close – the less frequently used sectors have the highest coefficients, the more heavily used have smaller coefficients.

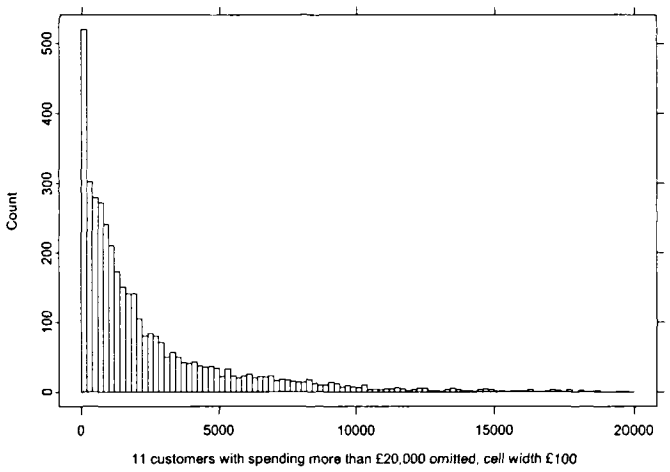
9.2.2 Discussion of this model

There are several features about these results that are unexpected.

Firstly, and ignoring the fact that the residuals are not normally distributed, or even close to it, is that supermarket spending should not – apparently – be included as an explanatory variable. Nor, probably, should 'other' or petrol. All of these are similar in one way – they are frequently used sectors. Why then, should 'other shop', which is similarly frequently used, be included – and as a negative impact – as its coefficients indicate? Similarly, why should the coefficients for clothing indicate that it should be excluded?

We show the distribution of customers' total spending in Figure 9.1. The cell widths are £200, and we have included the customers with no spending, but have omitted 12 customers with spending in excess of £20,000 (this is simply to avoid compressing the bulk of the data into the very left of the plot, a phenomenon we demonstrated in Chapter 8).

Figure 9.1 Total amount spent in a year



Although the people who spent these large amounts could be viewed as outliers, we consider that they are simply extreme and sparse values in a long tailed distribution. Unless our sample were to become very large, probably of the order of millions, there are always likely to be very few instances of such people. However, any modelling we undertake in the future will need to be able to accommodate such positively skewed distributions. Spending in most sectors, as we showed in Chapter 8, has distributions that are long tailed and right skewed. Note that this is rather different from the distribution of petrol station transactions we described in Chapter 7, where the vast majority of those were in a symmetric (or very close to symmetric) distribution, and values of more than £100 were clearly different in some qualitative way. Also, treating high values as outliers has another justification in that case – most cars have fuel tanks with a maximum capacity of around £50, so other transactions in that sector must (presumably) be for goods other than petrol. Also, if some are for lorry or coach fuel, those are also different from the majority, which are for personal car use.

Most distributions, such as transaction amount by sector, or spending by customer, are markedly right skewed, with a long tail. We are aware of only two other distributions in Barclaycard that are symmetrical, one of which describes a particular aspect of a specific type of correspondence, and the other a feature of telephone call handling times.

9.2.3 Transformations

In many instances, where data are skewed, log (or other) transformations are used, but we believe that they are not appropriate in this work, for two reasons.

In Table 8.1 we saw that even the most frequently used sectors were used by around half of our customers, which means that, should we want to log transform, we would have to add a constant to all values of each customer's spending in each sector. We would then have to decide what this might be, and it is not clear, although there are a number of possibilities, such as median spend, median spend in that sector, or 0.01. The latter would represent one penny and is the smallest amount that can be spent on a credit card.

There is a more fundamental problem, though, which is that most common transformations do not remove the skewness, but simply ameliorate the effect, or introduce skew in the opposite direction. Models may also become more difficult to interpret if we have a transformed value as the response variable.

9.2.4 Predicting one year's spend from the previous year

In Figure 9.2 we show the relationship between spending in two successive years (we have omitted two customers who spent more than £40,000 in 1996), and the diagonal line is a locally smoothed regression line. A simple linear model slightly under

predicts higher spending customers. It is obvious from this that there is a very good relationship between spending in 1996 and the following year; the correlation coefficient is 0.85. The most obvious conclusion from this is that to predict total spend at one time, we simply need to use earlier values of the same variable. Figure 9.3 shows the equivalent information for number, rather than value, of transactions; the correlation coefficient is 0.89. In this case, the local regression line virtually lies on the line $y = x$, so it would seem that, for spending, some other factor is coming into play that causes that relationship to be slightly weaker. We speculate that this could be inflation, as the effect is similar to the Retail Price Index in that year.

Figure 9.2 Total spend in two years

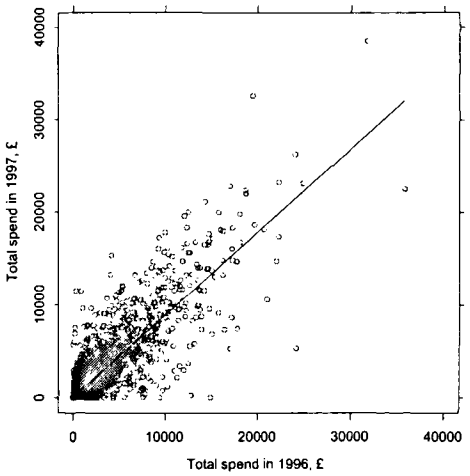
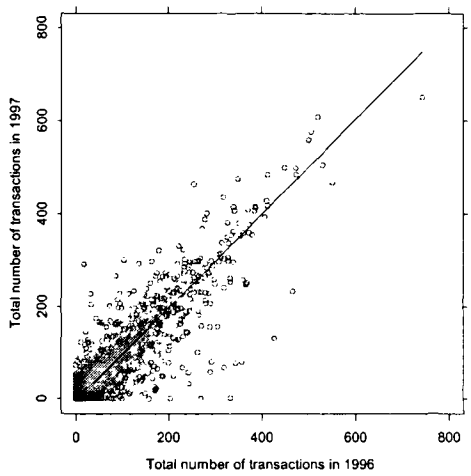


Figure 9.3 Number of transactions



A simple linear regression – with the outcome variable as the total amount spent per person in 1997, and only one predictor – total spend in 1996 – resulted in an R^2 of 0.72 (which we would expect, given the value of the correlation coefficient). More surprisingly, using spend in each of the 26 trade sectors as predictor variables, the adjusted R^2 only rose to 0.74. So, hardly any extra explanatory power is available by using information on all sectors.

The R^2 for the equivalent models using number of transactions rather than amounts spent were 0.80 and 0.81, for one predictor variable, and 26 individual sector predictors, respectively.

9.3 Predicting individual sectors' spend

9.3.1 Linear regressions

In Table 9.2 we show the R^2 statistic for a series of linear regressions, each of which used total spend per customer per sector in 1997 as the outcome variable. The left hand column shows the R^2 that resulted when we used only spending per customer

per sector in the *same* sector as the outcome variable. The centre column shows spending per customer in *all* sectors (i.e. a 3,972 × 26 matrix), and the right hand column is a 3,972 × 25 matrix, where the one variable omitted was the spending in the same sector. We adopted this approach to determine the different effects of (1) the same sector alone as predictor, (2) all 26 sectors together and (3) removing the same sector.

Table 9.2 Adjusted R^2 for linear regressions by sector

	Predictor variables		
Outcome sector	Same sector R^2	All sectors R^2	All, exc. same sector R^2
Cash advances	0.32	0.32	0.01
Chemist	0.47	0.51	0.26
Cinema	0.25	0.29	0.09
Clothing	0.52	0.57	0.36
CTN	0.47	0.48	0.06
Department store	0.45	0.51	0.29
DIY	0.23	0.30	0.20
Electrical goods	0.06	0.11	0.08
Furniture	0.03	0.10	0.09
Garage	0.25	0.30	0.30
Hotel	0.40	0.43	0.21
Jeweller	0.02	0.21	0.20
Mail order	0.15	0.22	0.15
Off licence	0.41	0.41	0.05
Other	0.47	0.54	0.24
Other finance	0.27	0.32	0.13
Other food	0.37	0.38	0.01
Other leisure	0.15	0.18	0.07
Other shop	0.44	0.48	0.33
Petrol stations	0.70	0.70	0.23
Public houses	0.05	0.10	0.07
Restaurant	0.64	0.66	0.35
Shoe shop	0.26	0.34	0.27
Supermarket	0.77	0.78	0.29
Transport	0.16	0.21	0.10
Travel agent	0.23	0.28	0.13

It is apparent from this that the most powerful predictor of spending in one sector is earlier values of spending in the same sector. It is also obvious that, for most sectors, adding behaviour in the 25 other sectors (as additional predictor variables) improves the predictions of the regressions by little. The only exceptions are jewellers and furniture stores, both of which are likely to be used spasmodically. For example, if a customer buys a three piece suite for, say, £1,000, she or he is unlikely to buy another one in the following year. To have a good prediction of spending in this sector, we may need many years of transaction history. Removing non-spenders from the models changed the R^2 statistics a little, but only by the order of 0.02.

9.3.2 Other predictor variables

We will not describe the results we obtained when we included other variables, such as repayment behaviour or (the limited) demographics we have in our data set. None of them had significant impacts on the results of the regressions. We discuss this more fully in Chapter 10.

9.3.3 Number of transactions

Using number, rather than value, of transactions as outcome and predictor variables makes little difference to the results of the models. Also, the predictive power of models built using number, rather than value, of transactions, is similar to those using the amounts spent. This is to be expected, given the roughly linear relationship between number and value of transactions, as we show for two sectors in Figure 9.4 and Figure 9.5 (the solid line is a local regression fit, and the dashed line is the linear regression fit).

Figure 9.4 Supermarket spend vs. number of transactions

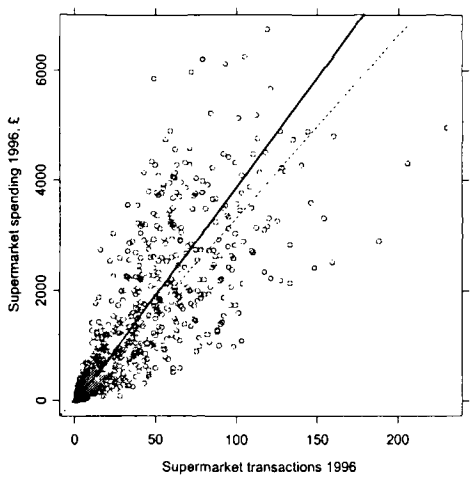
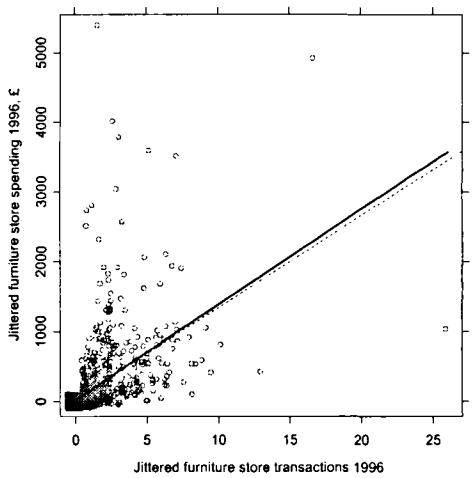


Figure 9.5 Furniture spend vs. number of transactions



The solid block at the origin is caused by the 79% of customers who did not use the sector in the course of a year. If we remove these customers the correspondence between the linear and local regression lines diverges a little more, but they are still close.

9.4 Conclusions

The best predictors of future transaction behaviour are earlier values of the same variable. The only exceptions to this are jewellers and furniture stores, where predictive power is markedly better if we include activity in all sectors. Using the data available to us, the simple solution to predict trade sector spending, by customer, in any sector, is to use earlier values of spending in the same sector. Further, to save data retrieval and analysis costs, it will not benefit the business to look at any other sectors. This is an important finding, because it means that costs can be limited by looking only at one sector when the aim is to predict spending in that sector. In a statistical sense, this might seem rather inadequate, and an obvious finding, but it gives considerable benefits in terms of better targeting and much greater response rates than are otherwise achievable. For commercially sensitive reasons we cannot say how much greater, only to note that it is definitely of benefit. This simplicity also means that we do not need to spend much (expensive) analysts' time on complex modelling, if a simple model produces good results.

There are only two exceptions to this rule, and, as we might have expected, they are sectors that tend to attract infrequent transactions. They are jewellers and furniture stores, sectors we might expect people not to use in two successive years.

Another refinement of the results described in this chapter, and for future work, is to look at, for example, spending in different airlines, rather than simply spending on air travel. This may be important, given that any incentive may well be with a limited subset of airlines, and it may be the business' aim in future to move customers' spending from one provider to another in the same sector.

However, as a guide for future work, there is benefit to continuing with research in new techniques and methods, because spending in different sectors has to be linked in some way, if only because most customers use a range of sectors. As computing power increases, and data sources become richer and more comprehensive, such work could reap rich rewards.

Chapter 10

10 Linking transaction and repayment behaviour

10.1 Introduction

We described, in Chapters 4 and 5, models for the prediction of different types of repayment behaviour, and in Chapter 9, for the prediction of spending behaviour. The most predictive variables in the first case were those on repayment behaviour, and in the second case those on spending behaviour.

As we described in Chapter 3, we have two broad types of data from these two different, but related, areas of customer behaviour, and the best linear models in each area never included variables from the other area as important predictors. The two types of behaviour *must* be related, because our data are taken from the same sample of customers and record their spending and borrowing activity. Expressing this slightly differently, if customer *A* spends this month, then she or he has to make a decision next month about how much of her or his balance to repay. At an individual level, therefore, it seems obvious that the two types of data have to be linked in some way. This chapter describes those links and how the business can make use of them, but it needs to use some non-linear techniques, or conditional analyses, to determine the links. They may not be strong (borne out by our investigations in earlier chapters) for the following reason.

Much qualitative consumer research is carried out at Barclaycard, and it gives some insight into customers' behaviour, which might indicate why only relatively weak quantitative links exist. We described, in Chapter 3, how credit cards are 'buy now,

pay later' products and this is borne out by the research. Consumers are able to dissociate the pleasurable – the purchase, and possession of the goods or services – from the less pleasurable, i.e. having to pay for their new possessions. These two events can be separated by almost two months – if customers buy at the beginning of their 'statement month' but do not make their repayment until the due date the following month. Furthermore, if the purchase and the payment are separate in many customers' minds, we should expect to find only weak links between the two different 'types' of data. If the customer is, or decides to become, a borrower the repayment could be spread over many months.

There is a relationship between spending and repayment behaviour, which we will describe, but also show how it is dwarfed by the 'spend-spend' and 'repayment-repayment' associations. The direction of the relationship is also what we expect, in that people who borrow more spend less, and vice versa. Using some simple measures, however, these relationships are not obvious, and further conditional analyses are necessary to deduce how they are related.

10.1.1 Segmentation versus clustering

Several authors draw a distinction between *clustering* and *segmentation*, for example Han and Kamber (2001), Hand, Mannila and Smyth (2001) and Berry and Linoff (2000). In each case, the authors use the term 'cluster analysis' in the same way that we did in Chapter 8 – the traditional statistical sense. It is a method of dividing n observations into g groups so that members of the same group are more alike than members of different groups, using some measure of dissimilarity. Segmentation, on the other hand, refers to a partition of the data that is convenient, and 'convenient' here refers to administrative convenience, practical convenience, or any kind that is

appropriate for the task. Often, when continuous data are partitioned in this way, it is with the objective of producing easy to read convenient or 'intuitive' partitions of the data. Take salary, for instance, which may be broken into a range such as (£20,000, £25,000) rather than (£19,883, £25,116).

We also draw a distinction between these two ways of dividing a data set. We show that, in our case, the relationships between spending and repayment behaviour are on a continuum, and that to use the relationships between them we need to choose two suitable partitions of the data. Again, 'suitability' depends on the nature of the task.

10.1.2 Amount spent and number of partial repayments

In Figure 10.1 we show the distribution of total spending compared to the number of partial repayments made. The latter are 'real partial repayments' as we defined them in Chapter 4, and both relate to activity in 1996. The relationship is obvious: the more partial repayments a customer makes, the less spending she or he makes. Not only this, but there is a marked drop in amounts spent for those customers making three or more partial repayments. The proportion of our sample making three or more partial repayments is 49.7% of the total. We show this figure for two reasons. Firstly, if we seek to segment the data, this might appear to be a suitable point, given the reduction in spending that we see among customers with three or more partial repayments. Secondly, if we are able to say to marketers 'half of the customers fall into each group' it is easy for non-numerate managers to understand the size of the groups. In Figure 10.2 we show the same information, but using the number of partial repayments this year and total spending the following year. It is almost identical to Figure 10.1, which should not be surprising, because spending one year

is highly correlated with spending the following year (we described this in Chapter 9).

Figure 10.1 Partial repayments and total amount spent in the same year

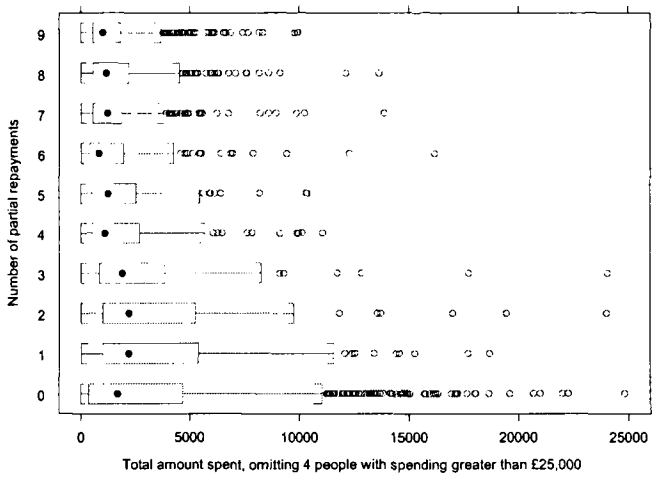
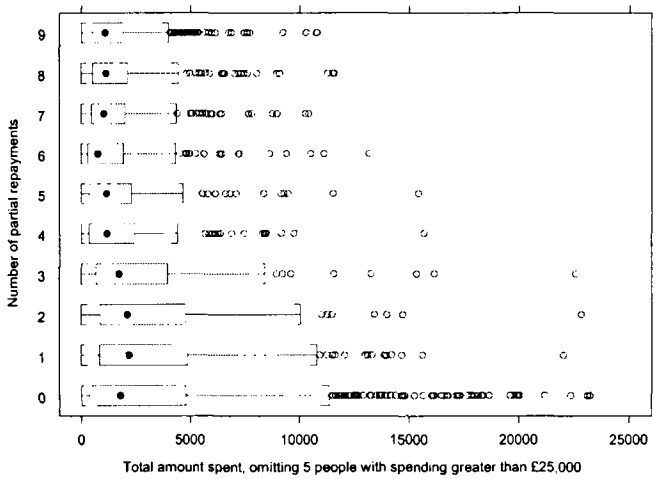


Figure 10.2 Partial repayments and total amount spent the following year



Similar investigations reveal relationships between the number of transactions made and the number of partial repayments, which should come as no surprise, because of the good relationship between the number of transactions made and amounts spent

on these transactions (see Chapter 8). Notice that slightly less is spent as the number of partial repayments reduces from one to none where we might have expected it to increase. The few customers (2.8%) who had no spending activity on their accounts cause most of this. If we remove them, we see that the median spend increases to almost the same as those people with one or two partial repayments.

10.2 Tree based models

10.2.1 Introduction

A classification or regression tree produces, respectively, a partition of the outcome space (of possible observations) into sub-regions corresponding to the leaves; or a constant prediction for each partition of the space corresponding to a leaf. For an explanation of classification and regression trees (CART) see Breiman et al. (1984), and for a description of C4.5 see Quinlan (1993). The method is non-linear, and uses recursive binary partitioning to produce groups ('leaves') that are successively more homogeneous, until it is infeasible to continue. This means that they may produce some insight in the present case, where we seek to examine the importance of one set of variables to another, where the relationship is not a simple linear one.

In order to assess the relationships between repayment and spending variables, we need to use more flexible methods than, say, linear regression, or other simple linear techniques. To demonstrate this, consider the following.

In Chapter 9 we described the use of linear regression to predict the amount spent by customers in each sector in 1997, from the amount spent in 1996, and saw an R^2 of 0.72. Using a simple linear regression to predict the number of transactions in 1997 using the number of transactions in 1996 as the sole predictor, we saw an R^2 statistic

of 0.80. Including the number of partial repayments in 1996 as a predictor variable as well, the R^2 statistic hardly improved, in fact by only 0.001. Using only the number of partial repayments as a predictor resulted in an R^2 statistic of only 0.05. (N.B. The reason for using number of transactions in this illustration, rather than their value, will become apparent).

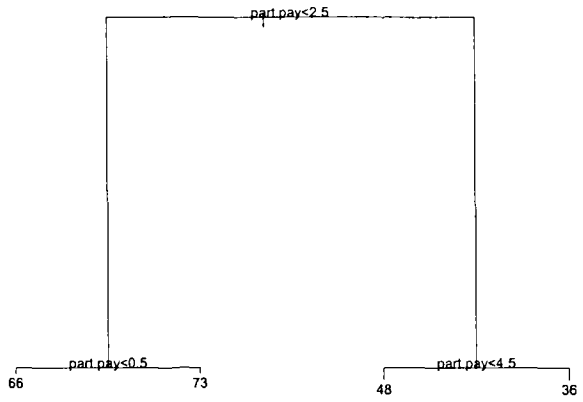
10.2.2 Number of transactions and repayment behaviour

In Figure 10.3 we show a tree developed using the number of transactions per year as the outcome variable, and the number of partial repayments in the year (as defined in Chapter 4) as the only predictor variable. At one level, it apparently works, because the predicted values at the nodes are in the order that we expected, given what we described in Section 10.1.2. However, notice that most of the predictive power comes from the first split; we have removed 6 leaves that added very little, based on the deviance of the full and pruned trees.

At this point, note that we chose to predict the number of transactions, rather than their value, because the trees are easier to interpret. This is simply because numbers of transactions are integers, and there are fewer of them – in our sample of 3,972 customers, there were more than 3,700 different amounts spent, but only 317 unique numbers of transactions. The results we obtained from trees where we used value of transactions as the outcome variable were similar to those using number of transactions, but the latter were easier to read, because of the smaller range of values.

The numbers at each leaf are the predicted numbers of transactions, and 'part.pay' is the variable showing the number of partial repayments. We used the change in deviance at each node as a guide to the vertical position of each pair of nodes.

Figure 10.3 Predicting the number of transactions from partial payments



This tree implies that, using only the number of partial repayments as a predictor, we would classify customers into four groups, with, respectively, 36, 48, 66 and 73 transactions in each one. This is (directionally) as we expected – in general, the more partial repayments, the fewer transactions. Also, the most significant split was between two and three partial repayments. Notice the same 'reversal' in the number of transactions when we reach '< 0.5' partial repayments (i.e. none), similar to the behaviour we showed in Figure 10.1. Again, the inactive customers cause this.

Predicting the number of transactions made in 1997 from partial repayments made in 1996 resulted in a model similar to the one shown in Figure 10.3.

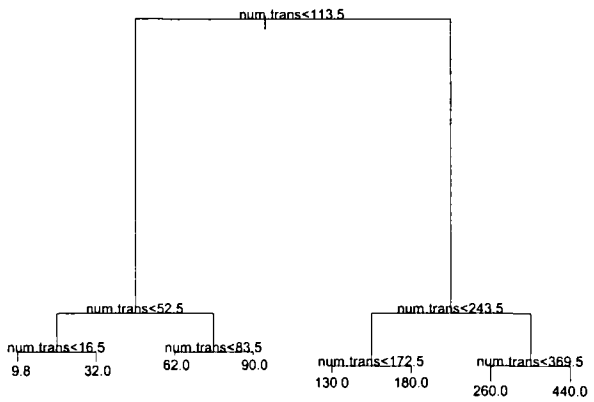
10.2.3 Predicting transactions the following year

In Figure 10.4, we show the results we obtained when predicting the number of transactions made in 1997 from repayment behaviour in 1996, and including the number of transactions made in 1996 as well (the labels 'num.trans' show the number of transactions made in 1996). As before, we have only shown those nodes

that improved the model, and notice now that partial repayments have disappeared (again using deviance change from node to node as our criterion for halting the tree).

The predicted range (of number of transactions) in each group is much broader in this model, indicating the groups are more homogeneous than those we saw in Figure 10.3, which used only repayment behaviour to predict number of transactions. This is as we expected, given the results of Chapter 9.

Figure 10.4 Predicting number of transactions from earlier values



10.2.4 Conclusions

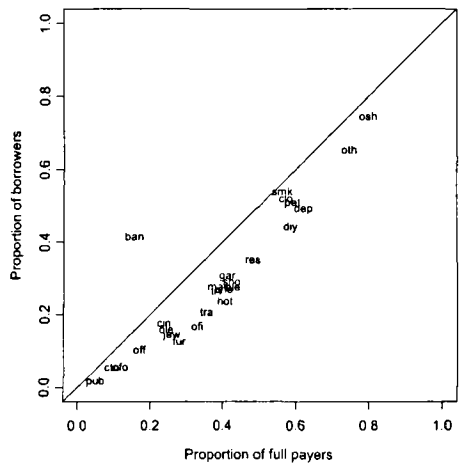
The conclusion to be drawn from this is the same as that from linear regression – although the number of partial repayments is related to the number of transactions, the relationship is weak, and it is overshadowed by transaction variables.

10.3 Differences between full payers and borrowers

10.3.1 Proportions using each sector

One area where we might expect differences between payers and borrowers is in the sorts of sectors they use. Qualitative research indicates that borrowers are much less likely to use supermarkets, because of the connotations that that might imply – having to borrow in order to buy food, for example. The data, however, indicate that the proportions of borrowers and full payers are similar for all sectors (with one exception). We show this in Figure 10.5, where the horizontal axis shows the proportion of full payers who use each sector, and the vertical shows the proportion of borrowers.

Figure 10.5 Proportions using each sector



In general, the differences are slight, as shown by the closeness of most sectors to the $y = x$ line. A slightly greater proportion of full payers (than borrowers) uses every sector, apart from cash advances ('ban'). For this analysis we have selected 'full payers' to be those who never made a partial repayment, and 'borrowers' as those

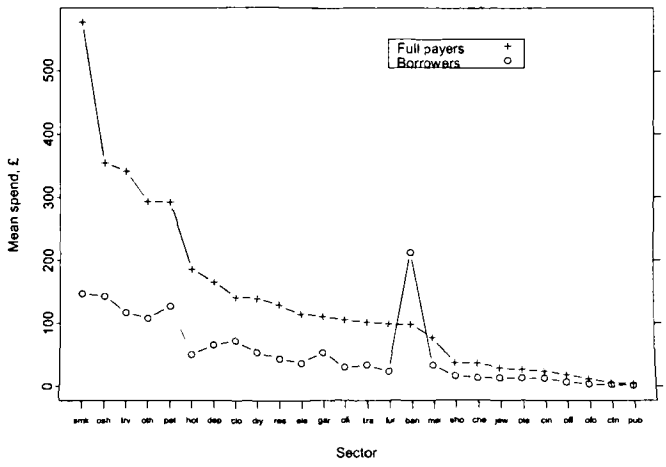
who always did. This ignores the people who sometimes pay and sometimes borrow, but the extreme groups comprise more than half of the sample, and any differences caused by repayment behaviour ought to be most distinct between these two extreme groups.

In fact, contrary to what customers tell us in research, supermarkets are used by similar proportions of borrowers and full payers. Customers do behave differently, but this is evidently by virtue of less frequent use in each sector, rather than the proportions using each sector.

10.3.2 Amounts spent by full payers and borrowers

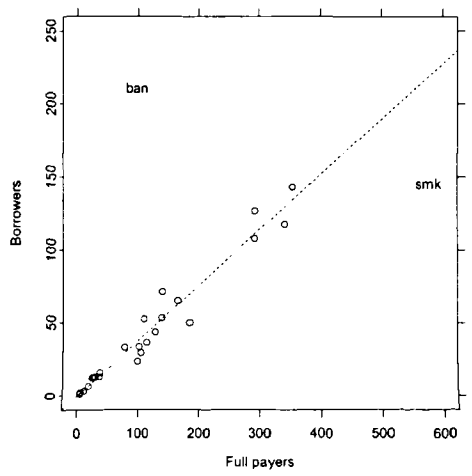
Despite these apparent similarities, there are large differences in the mean amounts that the two types of customer spend in each sector, which we illustrate in Figure 10.6. The horizontal axis represents each sector, and these are ordered by the mean amount spent by full payers. The '+' points show the mean spend per sector for full payers, and the 'o' points show the same for borrowers (for easier comparison, we have drawn lines between each consecutive pair of symbols). In most sectors, full payers spend have a higher mean spend than borrowers, and it is around 2.75 times as high. Once again, cash advances 'ban' – bank – are obviously different from other sectors. The sectors on the left of this plot tend to be the more heavily used ones, whereas the differences between the less heavily used sectors tend to be much less pronounced.

Figure 10.6 Mean spend per sector



In Figure 10.7, we show that the relationship between mean amount spent by payers and borrowers is close to linearity, with two exceptions – cash advances ('ban') and supermarkets ('smk'), which we have labelled, rather than showing them as points. The dashed line is the line of best fit, but excluding the two labelled points (note the different scales for the vertical and horizontal axes).

Figure 10.7 Payers and borrowers spend per sector



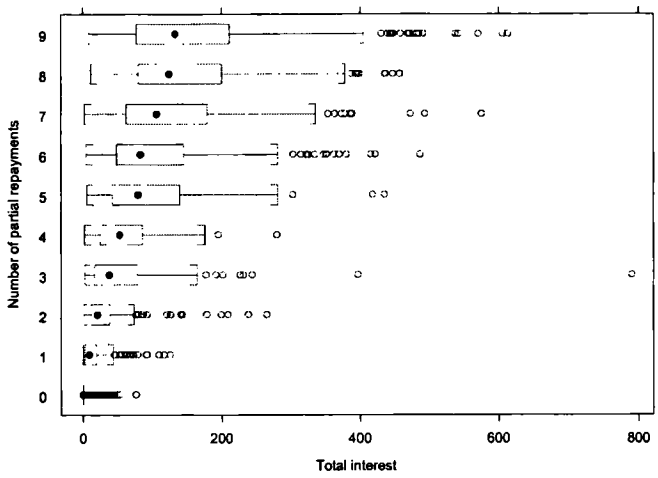
10.3.3 Interest and number of sectors used

There is a relationship between interest paid (over a year) and the number of partial repayments made, as we would expect, and we show this in Figure 10.8. Note that there are a small number of interest repayments made by people who were recorded with no partial repayments. We described, in Chapter 3, some examples of data distortion, and this is another one. Our definition of partial payers in Chapter 4 was quite specific, and was as follows.

$$x_{t,i} = \begin{cases} 0 \text{ if } \text{balance}_{t-1,i} \leq 0 \\ 1 \text{ if } \left(0 < \left(\text{payment}_{t,i} \div \text{balance}_{t-1,i}\right) < 1\right) \\ 0 \text{ otherwise} \end{cases}$$

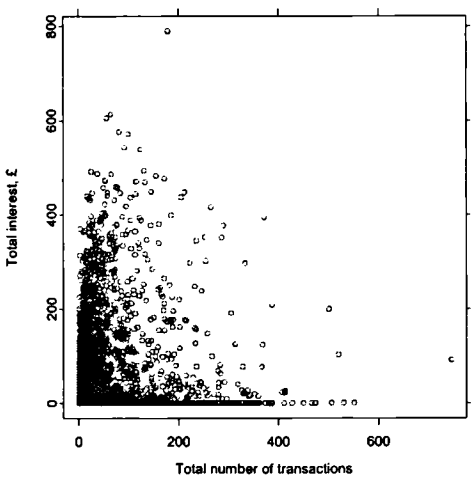
This excludes those people who miss their 'due date', but whose payment arrives before the next statement date. Statements are produced on the same billing day each month, and each customer has at least 25 days in which to make a payment, and there is always a gap of a few days between the due date and the next statement being produced. Payments arriving after the due date, but before the next statement date, will appear to have arrived on time, because they are recorded on the statement as having been made. If a payment is not received by the due date, however, interest will be charged on the balance outstanding on the statement. Thus, the account records that are held on file can show a full payment, but also that interest was charged. They make up a relatively small proportion of interest payments, and in some cases might be waived.

Figure 10.8 Interest and number of partial payments



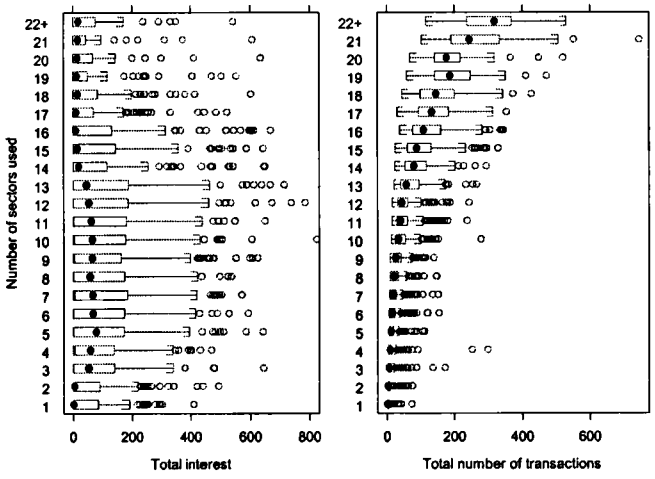
As the amount of interest is a good proxy for the value the business will realise from customers, it also makes sense to investigate the relationships between amounts spent (or number of transactions made) and interest paid. This is shown in Figure 10.9, and it is obvious that the relationship, if any, is highly non-linear, and probably not susceptible to being modelled by any sort of linear method. The correlation between these two variables is only -0.01.

Figure 10.9 Interest and transactions



However, there are some links between these two variables, and it becomes obvious when we introduce the number of sectors as well – see Figure 10.10. The more sectors used, the higher the spend, although this grows non-linearly, a phenomenon we noted in Chapter 8. We have omitted two points where customers had interest of more than £800 in the year. The amount of interest paid by those who spend in 15 or more sectors is considerably lower than that paid by people who use fewer sectors. The customers in the former account for 13.3% of interest, 50.0% of spending and 52.0% of transactions.

Figure 10.10 Including the number of sectors used



In Figure 10.11, we show how we can use this information for targeting, and it is by imposing a convenient segmentation. We have labelled the points corresponding to the number of sectors used (these are the same as the vertical axis in Figure 10.10). There are relationships between the number of transactions, the value of those, the number of sectors used and the amount of interest paid. They do not, however, cluster neatly if we use traditional clustering techniques – we have not shown these,

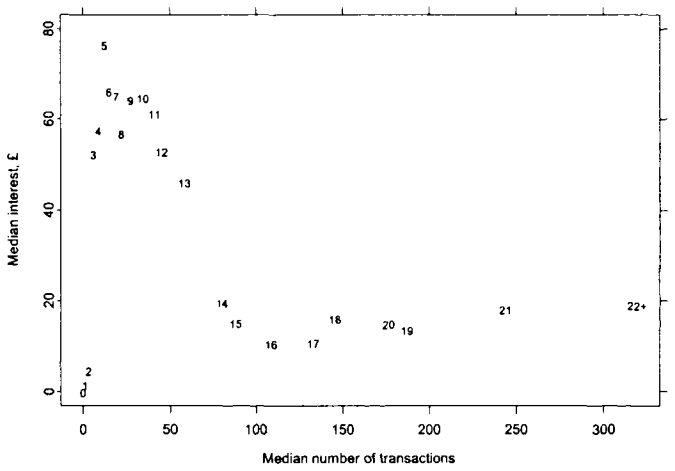
because the class separation is poor, and similar to the principal component plots we showed in Chapter 8.

We see, however, that, if we seek to maximise spending and interest paid, the most 'interesting' groups of customers are those who use between three and thirteen sectors, because they have the highest levels of interest. Also, almost two thirds of customers are in this group, which will give the business a large enough potential audience for targeting.

Customers who use most sectors, and make the greatest number of transactions, pay the least interest, and vice versa. There are two reasons why we might not target these customers. The first is that those with high spending are already using their cards for many items, possibly the majority of their spending; and they are may not be able to use it for much more. The second is that, if these customers pay little interest, it could take longer to make a return on our marketing investment. It is possible, but needs to be done with care – remember the example we cited in Chapter 2, in which one of the supermarket loyalty schemes was reported to be having problems. The scheme was reported to be costing the company many tens of millions of pounds, for little return.

Those with few transactions may well have the opportunity to use their card more, but the fact that they do not do so at the moment may indicate that there are some underlying reasons for this. Their behaviour may be difficult to shift, but for reasons different from those of the high spenders.

Figure 10.11 Number of transactions, sectors and interest



10.4 Conclusions

We have seen that there are relationships between different types of behaviour, but they do not become apparent from techniques such as cluster analysis. Thus, for use in a business context, and for targeting, we would impose our own segmentation, where we impose groupings for convenience, as distinct from clustering, where we seek to find natural groups, as a consequence of structures in the data.

We see that, if we select groups to incentivise, the customers we would choose to target are the 'moderate users'. We deduce this because – we suspect – we see a much higher proportion of an individual's spending among those customers using many sectors, and making a lot of transactions per year. There are two issues, but both with the same end result, that of encouraging more spending on our card; and of encouraging more people to move from other credit cards, or other methods of payment.

Almost as a side issue, by doing this, we will gain a little extra revenue, and potentially much more, if customers occasionally do not repay in full. We will measure this from shifts in the distributions in Figure 10.10, and the numbers of people using more sectors.

Chapter 11

11 Conclusions and future work

11.1 Conclusions

11.1.1 What we learned

We described two broad aspects of credit card use. The first was repayment behaviour, which concerned the way customers make repayments to their card accounts, how much of their balance they repay each month, and thus how much interest they are charged. We investigated how often partial repayments are made, and how often people are late with their payments. The second aspect described the ways that customers use their cards for spending, the relationships between different sectors, and how best to predict future use by sector, and the total amount spent.

We discovered that simple techniques (e.g. classical linear discriminant analysis, linear regression) performed better than more complex methods (e.g. k-nearest neighbour), when our a priori beliefs might have led us to believe otherwise. More importantly, the improvement we could achieve led to the discovery of target groups that contained up to three times as many suitable customers (for incentivisation) as we would expect if we were to take a random sample from the whole file. In the case of k-nearest neighbour, we found that classification performance was not only poorer at all values of k, but worsened markedly as we selected more neighbours. This could be improved by the use of a different threshold, but the results only matched those from LDA, they did not improve on them.

Visualising data is critical, because the maxim 'a picture tells a thousand words' can be crucially important in a business environment. It is not enough to undertake a robust piece of analysis; it must be presented in such a way that the conclusions and consequences are obvious to a non-technical audience. This is likely to be the senior management of the business, who are too busy to have the time to master some complex subjects in detail (and often their training will be in different disciplines). If this is done well, then funding for future initiatives is likely to be forthcoming; if done poorly, or insensitively, then funding may not. Also, if the analyst has a reputation for delivering high quality work, then it is easy to damage that high standing with one poor or ill considered piece of work.

This leads to conflicts, where often in a business environment the question to be answered is not 'is this model a good one?' (by whatever scoring criteria one adopts), but rather 'does it allow us to do things better than we used to?'. There are several reasons for this. Firstly, income pressures are always present, and it is often easier to assess short term gain than longer term retention of customers. We mentioned Reichheld (1993) in Chapter 2, who wrote 'at MBNA, a 5% increase in retention grows the company's profits by 60% by the fifth year'. It is far easier to measure incremental interest in a month's time than behaviour 60 months hence, and consequently, more effort tends to be devoted to measuring short term changes than longer term initiatives.

Secondly, in a business the size of Barclaycard, small percentage changes can have massive impacts. Thomas (1998) gives an example in which a reduction of bad debt by 0.25% could save a major issuer £20 million a year. A particular example arose in our modelling of petrol station rounders and non-rounders. By most statistical

assessments, we only found a weak predictor for the two groups of people, but that small difference was enough to make a substantial financial difference.

Different trade sectors showed markedly different characteristics of behaviour, although at individual transaction level the main reason was the pricing policies of the merchants concerned. There is scope for incentivising customers, and we showed that this could be targeted in such a way as to generate more income, by considering the relationships between spending and repayment (and thus the amount of income generated).

Credit card spending by sector can be modelled in the same way as branded consumer purchases, using the approach advocated by Ehrenberg (1959), Chatfield et al. (1966), Goodhardt et al. (1984) and Chatfield (1986). They fitted a negative binomial distribution to their purchasing data. For our data, the fit between observed and predicted values diverged at high numbers of transactions, in some sectors, which is similar to these authors' 'excessively heavy buyers' phenomenon. The NBD allowed us to parameterise, and simply, each sector, but it was of little use to predict spending, either between different sectors, or in the future. It did, however, form one of the measures in our taxonomy of distributions.

We considered four ways of modelling customers' use of different sectors, as follows.

- number of transactions per sector per customer
- transaction value in each sector
- number of sectors used
- conditional probabilities and odds ratios

The first, third and fourth of these are interesting from a business perspective, because increasing customers' use of their cards is a desirable objective for a credit card company. The second, although exhibiting interesting patterns, is of less use, because much of the structure is a consequence of pricing policies in different merchants.

We developed a taxonomy of spending. There are several ways to describe spending in different sectors and the differences between them, so we propose the following – which has two potential variations.

The first is transactions by sector, and groups sectors by customers' spending behaviour in each. There are sectors where transactions are rounded by choice (e.g. petrol stations), sectors where transactions are rounded by pricing policies (e.g. department stores), sectors where there is no rounding (e.g. supermarkets) and sectors where there are a small number of anomalous spikes (e.g. travel agents). Each of these fits into the mixture distribution framework we described in Chapter 8. It is also, of the types considered, the only method that might yield a discrete set of categories, but it requires some assumptions to be made about the degree of 'spikiness' in different sectors.

The second variation is to use one of three continua: the fit of the NBD to the number of transactions, or the relationships we deduced from conditional probabilities or odds ratios. We have to form a segmentation on these continua if we wish to create categories. We described the difference between *clustering* and *segmentation* in Chapter 10, and in this case, the segmentation we would choose would depend on the task in hand (e.g. which customers we were trying to incentivise in which trade sectors). The taxonomies we described are unlikely to provide a discrete set of

categories, but each will allow us to measure position along a continuum of various types of use. We would therefore seek to measure the success of any incentives or other marketing activity against the measures on each continuum.

Examples of the difficulties that can arise in data mining kept occurring. There were several instances when, by eye, models appeared to fit the data well, even though they failed significance tests. In most cases, we achieved predictions that were good enough to guide marketing activity. We described the way that customers are billed, which introduced its own distortions; the fact that later data extracts might not be consistent with earlier ones, although the same selection criteria might be used (and were, in our case). This is an inevitable fact of much data mining analysis – as more data and faster and bigger computers become available, systems are continuously upgraded, which improves data storage and retrieval at the time. If we were to commence work today, there would be more, and different, data fields available, and they might allow us to build better models. However, if we were to start collecting data now, it would be another two years before we would be able to start modelling, far less have any results.

11.1.2 What the business is doing differently

A powerful insight came from Chapter 4, where we were able to deduce the proportion of customers who would always make a partial repayment, those who would never do so, and the proportions at all levels in between the two extremes. This has provided strategic guidance for the business, because of the knowledge about the proportions of people who are likely to remain full payers. Interest income is an important component of credit card profitability, as we described in Chapter 2,

but Barclaycard has a substantial proportion of customers who seem unlikely – ever – to pay much interest.

The business therefore needs to find ways of offering other, or different, products and services to such customers. There are two reasons for this: firstly, full payers are less likely to be concerned about interest rates, but are more likely to seek an incentive related to spending. We described the advantages and disadvantages of such schemes in Chapter 2, where we noted that one retailer has done very well, and others poorly, from their loyalty schemes.

Barclaycard now targets different groups of customers with a variety of incentives that were not offered in the past, based on their spending behaviour. It is important to be able to select suitable groups for targeting, but it is equally critical to know which customers to target for the likely best returns, based on a combination of their repayment and purchasing patterns.

11.2 Future work

These are suggestions for work that could be done in the future, and they follow from questions that have arisen during the course of this work.

11.2.1 Repayment behaviour

We did not examine how patterns of repayment behaviour were changing. Partly this was because of the stationary nature of the data we had available, so we would need to consider a longer period than only two years. We would seek to develop measures of atypicality, and find ways to identify and predict when a customer's behaviour will change.

We should consider population drift – our sample hardly suffered from this, because it only spanned two years, but in Chapter 1 we described how Barclaycard's product composition since has changed since our data were extracted. A related issue is that of generalising from a shrinking sample. One possibility would be the use of weighted analysis to take into account the greater confidence in the cases with more observations.

Customers who regularly miss payment dates might be different from those who never do. Our work simply predicted future behaviour, but with extra data we should be able to model how different these groups are, and what alternative products and services we could provide that might be more suitable.

When using LDA to classify repayment behaviour, we ignored the ordinal information implicit in the data. It would be of value to find out how much extra predictive power this might give us. However, new methods may have to be developed to do this – see for example, Hand, Li and Adams (2001).

We allocated a value of zero to customers with zero and negative balances. This worked reasonably well for our purposes, but there may well be a better way to include these values in our modelling. Also, for prediction of repayment behaviour, we used a 'cut-off' value of 100% of balance repaid, but a cut-off at a different level might have performed better.

In the k-nearest neighbour work, we should consider more approaches for choosing the classification threshold. Firstly, in the two class case, an alternative way of choosing the threshold would be to optimise directly the criterion of interest (e.g. the difference between the two ratios, $d/(c+d)$ and $D/(C+D)$). Secondly, we should

investigate the relationship between k, the threshold t, and costs. Thirdly, in the 10 class case, we should consider different ways of allocating class membership amongst the k nearest neighbours. This would be instead of the simple choice of the class that has the maximum points amongst the k nearest neighbours, and it would also need to incorporate the costs of misclassification.

11.2.2 Spending behaviour

We might consider how the timing of transactions, and their frequency and amount varies across different groups of customers. As part of this, we should examine separately the highest spending 2% - 3% of customers. In order to accomplish this, we need to characterise features of behaviour such as alternating bursts of rapid spending and no spending, or long breaks punctuated by occasional large spends.

Transactional data might offer greater predictive power if we used it in a less aggregated way. The obvious choice would be to use individual Merchant Category Codes (MCC), although this would give rise to problems as well. As we described in Chapter 7, the number of MCCs can be quite different in different categories – for example supermarkets, which are all contained in two codes, compared to (approximately) 300 for airlines.

Some sectors (e.g. hotels) have transactions for 1p, or similar small amounts. It might be possible to develop an automatic way of screening for these transactions. As described in Hand, Blunt, Kelly and Adams (2000), automatic fault rectification can mask the very patterns we are seeking. Transactions for exactly 1p are easy to find, but DIY seems prone to having a small proportion that have values of a few

pence, and for different amounts. Identifying the genuine ones would not be easy in this case.

What are the effects, and how should we include them, of customers with no spending in particular months? In our analyses we gave them the value zero, but we suspect the 'never spend', the 'occasionally spend' and the 'always spend' customers are not the same. We should model the differences between them, and more importantly, develop models to predict when people will stop using their cards. We would also need to assess over what period should they be considered as non-spenders.

We showed, in Chapter 8, that those who used their card in more sectors spent disproportionately more per sector. This curve and other analyses suggest that a useful measure of behaviour is captured by the notion of 'propensity to use the card'. We should develop a formal modelling methodology for this concept.

11.2.3 General

Gold, Platinum and other types of new accounts have been launched in recent years, and, in general, there has been much movement of customers to new products. At the time of our data extract, Barclaycard's standard Visa file amounted to 90% of the total business, but has fallen to around half today. More and more new products have been launched, all targeted at much smaller segments of the population. We should investigate whether these different products have changed customers' behaviour.

Much richer sources of data are now available, compared to the time we started this work. They would enable us to partition the population by other types of variable,

and possibly identify which sub-populations can be predicted with high accuracy. A subsequent question would be to ask if we are using the best predictor variables, and to investigate whether others exist that would enhance the power of what we are seeking to do. Also, as computing technology has developed, we can deal more easily with much larger data sets than we could a few years ago.

We removed a substantial number of accounts from our sample. This could be important because 'lost & stolen' accounts tend to be more active than the average, and future modelling should include them. Similarly, we omitted new customers. However, we would need a bespoke sample of customers to model the behaviour of people in this group. We would seek to model characteristics such as time to first use, time to first borrowing, balance growth in the early months, and the relationship between the sectors used first and the propensity to pay interest.

Many of our variables are on very different scales (e.g. transaction amount and interest paid), and these might affect the distance measures we use in techniques such as *k*-nearest neighbour analysis. We need to have a rigorous framework for assessing what the optimal metrics might be.

It might be possible to construct models that enable us to predict which customers we can predict accurately into a particular classification schema. This would possibly be the most powerful insight we could ever develop: a measure of uncertainty in the whole customer base.

Bibliography

Adderley R. and Musgrove P. B. (2000), Data mining at the West Midlands Police: a study of bogus official burglaries, *Proc. 19th SGES International Conference On Knowledge-Based Systems And Applied Artificial Intelligence*, 191-203.

Adams, R. M. and Hand, D. J.. (1999), Comparing classifiers when the misallocation costs are uncertain, *Pattern Recognition*, **32,** 1139-1147.

Agrawal, R., Imielinski, T. and Swami, A. (1993a), Mining association rules between sets of items in large databases, *Proc. of the 1993 ACM SIGMOD Conference on the Management of Data*, 207-216, Washington D.C.

Agrawal, R., Imielinski, T. and Swami, A. (1993b), Database mining: a performance perspective, *IEEE Transactions on knowledge and data engineering*, **5,** 914-925.

Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *Proc. of the 20th Conference on Very Large Databases*, 478-499.

Allenby, G. M., Arora, N. and Ginter, J. L. (1998), On the heterogeneity of demand, *Journal of Marketing Research,* **XXXV**, 384-389.

Allenby, G. M. and Ginter, J. L. (1995), Using extremes to design products and segment markets, *Journal of Marketing Research*, **XXXII**, 392-403.

Allenby, G. M., Leone, R. P. and Jen, L. (1999), A dynamic model of purchase timing with application to direct marketing, *Journal of the American Statistical Association*, **94**, 365-374.

Allenby, G. M. and Lenk, P. J., (1994), Modeling household purchase behavior with logistic normal regression, *Journal of the American Statistical Association*, **89**, 1218-1231.

Allenby, G. M. and Rossi, P. E. (1999), Marketing models of consumer heterogeneity, *Journal of Econometrics*, **89**, 57-78.

APACS (2001a), *Card fraud - the facts*, Association for Payment Clearing Services, London.

APACS (2001b), *Plastic Card Review*, Association for Payment Clearing Services, London.

APACS (2001c), *Consumer Payments Survey*, Association for Payment Clearing Services, London.

Auriemma, M. J. and Coley, R. S. (1992), *The Bankcard Business*, The American Bankers Association, United States of America.

Ausubel, L. M. (1997), Credit card defaults, credit card profits, and bankruptcy, *American Bankruptcy Law Journal*, **71**, 249-270.

Bank of England (1998), *Lending to Individuals*, November, Bank of England, London.

BBA (2001), *Annual Abstract of Banking Statistics*, **18**, British Banker's Association, London.

Berry, M. J. A. and Linoff, G. (2000) *Mastering Data Mining. The Art and Science of Customer Relationship Management*, Wiley, New York.

Berthoud, R. and Kempson, E. (1992), *Credit and Debt: the PSI Report,* Policy Studies Institute, London.

Bhargava, A. and Sargan, J. D. (1983), Estimating dynamic random effects models from panel data covering short periods, *Econometrica*, **51**, 1635-1659.

Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities. *Journal of Political Economics*, **81**, 637-659.

Bolton, R. J. and Hand D. J. (2001), Unsupervised Profiling Methods for Fraud Detection, *Proc. Credit Scoring and Credit Control VII*, Edinburgh.

Box, G. E. P., Jenkins G. M. , and Reinsel G. C. (1994), *Time series analysis: forecasting and control*, Prentice Hall, Englewood Cliffs, New Jersey.

Boxall, G. (1996), *New Card Technologies in Retail Banking: Competition and Collaboration in the 1990s,* MPhil thesis, The Open University.

Boulicaut, J-F., Klemettinen, M. and Mannila, H. (1998), Querying inductive databases: a case study on the MINE RULE operator, *Lecture Notes in Artificial Intelligence*, **1510**, 194-202.

Bradley, P. S., Fayyad, U. M. and Mangasarian, O. L. (1999), Mathematical programming for data mining: formulations and challenges, *Informs Journal on Computing*, **11**, 217-238.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont California: Wadsworth.

Brito, D. L. and Hartley, P. R. (1995), Consumer rationality and credit cards, *The Journal of Political Economy*, **103**, 400-433.

British Market Research Bureau (2001), *Target Group Index*, British Market Research Group: London.

Brooking, C. (1997), The UK credit card market, *The Bank of England Financial Stability Review*, Autumn, 92-104.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, J. W. (1983), *Graphical Methods for Data Analysis,* Wadsworth, California.

Chan, P. K., Fan, W., Prodromidis, A. L. and Stolfo, S. J. (1999), Distributed data mining in credit card fraud detection, *IEEE Intelligent Systems & Their Applications*, **14**, 67-74.

Chiang, J., Chung, C-F. and Cremers, E. T. (2001), Promotions and the pattern of grocery shopping time, *Journal of Applied Statistics*, **28**, 801-819.

Chatfield, C. (1986), Discrete distributions and purchasing models, *Communications in Statistics*, **15,** 697-708.

Chatfield, C. (1995), Model uncertainty, data mining and statistical inference, *Journal of the Royal Statistical Society Series A*, **158,** 419-466.

Chatfield, C., Ehrenberg, A. S. C. and Goodhart, G. J. (1966), Progress on a simplified model of stationary purchasing behaviour, *Journal of the Royal Statistical Society Series A*, **124,** 326-367.

Cleveland, W. S. (1985), *The Elements of Graphing Data,* Wandsworth advanced books and software, Monterey, California.

Cleveland, W. S. (1993), *Visualizing Data,* Hobart Press, New Jersey.

CCRG (2001), *Statistical Yearbook*, Credit Card Research Group: London.

Cox C. K., Eick S. G., Wills, G. J. and Brachman, R. J. (1997), Visual data mining: recognizing telephone calling fraud, *Data Mining And Knowledge Discovery*, **1**, 225-231.

Crook J. N., Thomas L. C. and Hamilton R. (1994), Credit cards: haves, have-nots and cannot-haves, *The Service Industries Journal*, **14**, 204-215.

Cruickshank, D. (2000), *Competition in UK Banking: A Report to the Chancellor of the Exchequer*, The Stationery Office, Norwich.

Cyert, R. M., Davidson, H. J. and Thompson, G. L. (1962), Estimation of the allowance for doubtful accounts, *Management Science*, **8**, 287-303.

Davies, R. B. and Pickles, A. R. (1987), A joint trip timing store-type choice model for grocery shopping, including inventory effects and nonparametric control for omitted variables, *Transportation Research*, **21A**, 345-361.

Dekimpe, M. G., Hanssens, D. M. and Silva-Risso, J. M. (1999), Long-run effects of price promotions in scanner markets, *Journal of Econometrics*, **89**, 269-291.

Devijver P. A. and Kittler J. (1982), *Pattern Recognition: a Statistical Approach.* Englewood Cliffs, New Jersey: Prentice Hall.

Dong, G. and Li, J. (1997), Interestingness of discovered association rules in terms of neighborhood-based unexpectedness, *Technical Report 97/24*, University of Melbourne.

DunnHumby Ltd., (2001), *Tesco: a case study*, DunnHumby Ltd.: London.

Edwards, D. (1995), *Introduction to Graphical Modelling*, Springer Verlag, New York.

Efron, B. and Tibshirani R. J. (1993), *An Introduction to the Bootstrap,* Chapman and Hall, Monographs on Statistics and Applied Probability, London.

Ehrenberg, A. S. C. (1959), The pattern of consumer purchases, *Applied Statistics*, **8**, 26-41.

Elder J. and Pregibon D. (1996), A Statistical Perspective on KDD, *Advances in Knowledge Discovery and Data Mining, Eds. Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.*, AAAI/MIT Press, 83-116.

Fayyad, U. M., Djorgovski, S. G. and Weir, N. (1996a), From digitized images to online catalogs - data mining a sky survey, *AI Magazine*, **17,** 51-66.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996b), From data mining to knowledge discovery in databases: an overview, *Advances in Knowledge Discovery and Data Mining, Eds. Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.*, AAAI/MIT Press, 1-34.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996c), From data mining to knowledge discovery in databases, *AI Magazine*, **17**, 37-54.

Feinberg, R. A. (1986), Credit cards as spending facilitating stimuli: a conditioning interpretation, *Journal of Consumer Research*, **13**, 348-356.

Foekens, E. W., Leeflang, P. S. H. and Wittink, D. R. (1999), Varying parameter models to accommodate dynamic promotion effects, *Journal of Econometrics*, **89**, 249-268.

Friedman J. H. (1997), Data mining and statistics: what's the connection?, *Proc. 29th Symposium on the Interface Between Computing Science and Statistics*, 3-9.

Freitas A. A. (1998), A multi-criteria approach for the evaluation of rule interestingness, *Proc. International Conference On Data Mining*, Brazil, 7-20.

Freitas A. A. (1999), On rule interestingness measures, *Knowledge-Based Systems*, **12**, 309-315.

Frydman, H., Kallberg, J. and Kao, D. (1985), Testing the adequacy of Markov Chain and Mover-Stayer models as representations of credit behaviour, *Operations Research*, **33**, 203-1214.

Fuller W. A. (1996), *Introduction to statistical time series*, John Wiley and Sons, New York.

Fukunaga, K. and Flick, T.E. (1984), An optimal global nearest neighbour metric, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 25-37.

Ganzach, Y. and Karsahi, N. (1995), Message framing and buying behavior: a field experiment, *Journal of Business Research*, **32**, 11-17.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1996), Statistical inference and data mining, *Communications of the ACM*, **39**, 35-41.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1997), Statistical themes and lessons for data mining, *Data Mining and Knowledge Discovery*, **1**, 11-28.

Goodall, C. R. (1999), Data mining of massive datasets in healthcare, *Journal of Computational And Graphical Statistics*, **8**, 620-634.

Goodhardt, G. J., Ehrenberg, A. S. C. and Chatfield, C. (1984), The Dirichlet: a comprehensive model of buying behaviour, *Journal of the Royal Statistical Society Series A*, **147**, 621-655.

Greenfield, A. (1994), Comment on deconstructing statistical questions (Hand, 1994), *Journal of the Royal Statistical Society Series A*, **157**, 339-341.

Ha, S. H. and Park, S. C. (1998), Application of data mining tools to hotel data mart on the intranet for database marketing, *Expert Systems With Applications*, **15**, 1-31.

Han, J. and Fu, Y. (1999), Mining multiple-level association rules in large databases, *IEEE Transactions on knowledge and data engineering*, **11**, 1-8.

Han J. and Kamber M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco.

Hand, D. J. (1981), *Discrimination and Classification*, Wiley, New York.

Hand D. J. (1994), Deconstructing statistical questions (with discussion), *Journal of the Royal Statistical Society Series A*, **157**, 317-356.

Hand, D. J. (1997), *Construction and Assessment of Classification Rules*, Wiley, New York.

Hand D.J. (1998a), Data mining: statistics and more? *The American Statistician*, **52**, 112-118.

Hand, D.J. (1998b), Consumer credit and statistics. In *Statistics in Finance*, ed. D.J. Hand and S.D. Jacka. London, Arnold, 69-81.

Hand D. J. (2000a), Methodological issues in data mining, *COMPSTAT 2000: Proceedings in Computational Statistics*, ed. J.G. Bethlehem and P.G.M. van der Heijden. Physica-Verlag, 77-85.

Hand D. J. (2000b), Data mining - new challenges for statisticians, *Social Science Computer Review*, **18**, 442-449.

Hand D. J. and Blunt, G. (2001), Prospecting for gems in credit card data, *IMA Journal of Mathematics Applied in Business and Industry*, **12**, 173-200.

Hand D. J., Blunt, G. and Bolton, R. J. (2001), A note on confidence and support, *Technical Report*, Department of Mathematics, Imperial College, London.

Hand D. J., Blunt, G., Kelly, M. G. and Adams, N. M. (2000), Data mining for fun and profit, *Statistical Science*, **15**, 111-131 (with discussion).

Hand, D. J. and Henley, W. E. (1997) Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society Series A*, **160**, 523-541.

Hand, D. J. and Jacka, S. D. (Eds.) (1998), *Statistics in Finance*, Arnold Applications of Statistics, London.

Hand, D.J., Li, H. G. and Adams, N. M. (2001), Supervised classification with structured class definitions, *Computational Statistics and Data Analysis*, **36**, 209-225.

Hand D.J., Mannila H. and Smyth P. (2001b), *Principles of Data Mining*, MIT Press.

Hand, D.J., McConway, K.J. and Stanghellini, E. (1997), Graphical models of applicants for credit. *IMA Journal of Mathematics Applied in Business and Industry*, **8**, 143-155.

Hand, D.J., Oliver, J. J. and Lunn, A. D. (1998), Discriminant analysis when the classes arise from a continuum. *Pattern Recognition*, **31**, 641-650.

Harvey A. C. (1989), *Forecasting, structural time series models, and the Kalman filter*, Cambridge University Press, Cambridge .

Heckerman, D. (1997), Bayesian networks for data mining, *Data Mining and Knowledge Discovery*, **1**, 79-119.

Henley, W. E. and Hand, D. J. (1996), A k-nearest-neighbour classifier for assessing consumer credit risk, *The Statistician*, **45**, 77-95.

Henley, W. E. and Hand, D. J. (1997), Construction of a k-nearest-neighbour credit scoring system, *IMA Journal of Mathematics Applied in Business and Industry*, **8**, 305-321.

Hipp, J., Guntzer, U. and Nakhaeizadeh, G. (2000), Algorithms for association rule mining – a general survey and comparison, *SIGKDD Explorations*, **2**, 58-64, ACM.

Hoadley, B. and Oliver, R. M. (1998), Business measures of scorecard benefit, *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 55-64.

Hofmann, H. and Wilhelm, A. (2001), Visual comparison of association rules, *Computational Statistics*, **16**, 399-415.

Huber, P. J. (1997), From large to huge: a statistician's reactions to KDD & DM, *Proc. of the Third International Conference on Knowledge Discovery and Data Mining*, 304-308.

Hunziker, P., Maier, A., Nippe, A., Tresch, M., Weers, D. and Zemp, P (1998), Data mining at a major bank: lessons from a large marketing application, *Lecture Notes In Artificial Intelligence*, **1510**, 345-351.

Jha, G. and Hui, S. C. (1998), Data mining for risk analysis and targeted marketing, *Lecture notes in Artificial Intelligence*, **1531**, 158-169.

Kallberg, J. G. and Saunders, A. (1983), Markov Chain approaches to the analysis of payment behaviour of retail credit customers, *Financial Management*, **12**, 5-14.

Kamakura, W. A. and Russell, G. J. (1989), A probabilistic choice model for market segmentation and elasticity structure, *Journal of Marketing Research*, **XXXIII**, 379-390.

Keim, D.A. and Kriegel, H. P. (1996), Visualization techniques for mining large databases: a comparison, *IEEE Transactions on Knowledge and Data Engineering*, **8,** 6, 923-938.

Kemp, R., Towell, N. and Pike, G. (1997), When seeing should not be believing: photographs, credit cards and fraud, *Applied Cognitive Psychology*, **11**, 211-222.

Klemettinen, M., Mannila, H. and Toivonen, H. (1997), A data mining methodology and its application to semi-automatic knowledge acquisition, *Proc. Eighth International Workshop on Database and Expert Systems Applications*, 670-8677.

Leszczyc, P. T. L. P. and Bass, F. M. (1998), Determining the effects of observed and unobserved heterogeneity on consumer brand choice, *Applied stochastic models and data analysis*, **14**, 95-115.

Macedo, M., Cook, D., B. and Timothy J. (2000), Visual Data Mining in Atmospheric Science Data, *Data Mining and Knowledge Discovery*, **4**, 69-80.

MacKinnon, M. J. and Glick, N. (1999), Data mining and knowledge discovery in databases – an overview, *Australian and New Zealand Journal of Statistics*, **41**, 255-275.

Mannila, H. (1996), Data mining: machine learning, statistics, and databases, *Proc. Eighth International Conference on Scientific And Statistical Database Systems*, 2-9.

Mannila, H. (1997), Methods and Problems in Data Mining, *Lecture Notes in Computer Science*, **1186**, 41-55.

Mannila, H., Toivonen, H., Korhola, A. and Olander, H. (1998), Learning, mining or modeling? A case study from paleoecology, *Proc. First International Conference on Discovery Science*, 12-24.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.

McLachlan G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

AC Nielsen MMS Market Movements (2001), *Direct Mail Monitor*, Royston, Hertfordshire.

MoneyFacts (2002), *MoneyFacts ONLINE*, Norwich.

The Monopolies and Mergers Commission (1989), *Credit Card Services*, The Stationery Office, Norwich.

NOP World (2001), *Financial Research Survey*, London.

Padmanabhan B. and Tuzhilin A. (1999), Unexpectedness as a measure of interestingness in knowledge discovery, *Decision Support Systems*, **27**, 303-318.

Park, S. (1997), Effects of price competition in the credit card industry, *Economics Letters*, **57**, 79-85.

Parker, G. (1990), *Getting and Spending. Credit and Debt in Britain*, Avebury, Aldershot.

Parzen E. (1997) Data mining, statistical methods mining, and history of statistics, *Mining and modeling massive data sets in science, engineering, and business with a subtheme in environmental statistics 29(1): Computing Science and Statistics (series), 365-374*, 1997.

Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. (1999), Efficient mining of association rules using closed itemset lattices, *Information Systems*, **24,** 25-46.

Piatetsky-Shapiro, G., Matheus, C., Smyth, P. and Uthurusamy, R. (1994), KDD-93, Progress and Challenges in Knowledge Discovery In Databases, *AI Magazine*, **15,** 77-82.

Pregibon D. (2000), Signature-based methods in data mining, *Proc. of the Workshop on Statistical Modelling for Data Mining, University of Pavia,* October 25th - 26th.

Queen, C. M. (1999), Using the multiregression dynamic model to forecast brand sales in a competitive product market, *The Statistician*, **43**, 187-98.

Quinlan J. R. (1993), *C4.5: Programs for machine learning*, San Mateo, CA: Morgan Kaufmann.

Reichheld, F. F. (1993), Loyalty-based management, *Harvard Business Review*, **71**, 64-73.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Ritzer, G. (1995), *Expressing America: a critique of the global credit card society,* Pine Forge Press, California.

Romaniuk, H., Skinner, C. J. and Cooper, P. J (1999), Modelling consumers'use of products, *Journal of the Royal Statistical Society Series A*, **162,** 407-421.

Rosenberg, E. and Gleit, A. (1994), Quantitative methods in credit management: a survey, *Operations Research*, **42**, 589-613.

Rud, O. P. (2001), *Data Mining Cookbook*, Wiley, New York.

Sahar, S. (1999), Interestingness via what is not interesting, *Proc. of the Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 332-336.

Smyth, P. (2000), Data mining: data analysis on a grand scale?, *Statistical Methods in Medical Research*, **9**, 309-327.

Simpson C. H. (1951), The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society Series B*, **13**, 238-241.

Stanghellini, E., McConway, K.J. and Hand, D.J. (1999) A discrete chain graph for applicants for credit. *Journal of the Royal Statistical Society Series C*, **48**, 239-251.

Stavins, J. (1996), Can demand elasticities explain sticky credit card rates?, *New England Economic Review*, **July/August**, 43-54.

Office for National Statistics (2001), *Annual Abstract of Statistics*, 137. The Stationery Office, Norwich.

Thomas, L. C. (1998), Methodologies for classifying applicants for credit. In *Statistics in Finance*, ed. D.J. Hand and S.D.Jacka., Arnold, London., 83-103.

Till, R. (2001), *Predictive behavioural models in credit scoring and retail banking*, Unpublished PhD thesis, Department of Mathematics, Imperial College, London.

Tukey J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Webb A. (1999), *Statistical Pattern Recognition*, Arnold, London.

Wegman, E. (1995), Huge data sets and the frontiers of computational feasibility, *Journal of Computational and Graphical Statistics*, **4**, 281-295.

Witten, I. H. and Frank, E. (2000), *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufman, San Francisco.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics,* Wiley, New York.

Wrigley, N. and Dunn, R. (1985), Stochastic panel-data models of urban shopping behaviour: 4. Incorporating independent variables into the NBD and Dirichlet models, *Environment and Planning A*, **17,** 319-331.